Journal of Machine Learning Research 26 (2025) 1-37 Submitted 6/20; Revised 3/25; Published 10/25

Kernel-based L_2-Boosting with Structure Constraints

Yao Wang YAO.S.WANG@GMAIL.COM
Center for Intelligent Decision-Making and Machine Learning
School of Management
Xi'an Jiaotong University, Xi'an, China
Xin Guo XIN.GUO@UQ.EDU.AU
School of Mathematics and Physics
The University of Queensland
Brisbane, QLD 4072, Australia
Shao-Bo Lin* SBLIN1983@GMAIL.COM
Center for Intelligent Decision-Making and Machine Learning
School of Management
Xi'an Jiaotong University, Xi'an, China

Editor: Ingo Steinwart

Abstract

Developing efficient kernel methods for regression is popular in the past two decades. In this paper, utilizing boosting on kernel-based weak learners, we propose a novel kernel-based learning algorithm called kernel-based re-scaled boosting with truncation, dubbed as KReBooT. The proposed KReBooT benefits in controlling the structure and producing sparse estimators, and is near overfitting resistant. We conduct both theoretical analysis and numerical simulations to illustrate the excellent performance of KReBooT. Theoretically, we prove that KReBooT can achieve the optimal numerical convergence rate for nonlinear approximation. Furthermore, using a variant of Talagrand's concentration inequality, we provide fast learning rates for KReBooT, which is a new record of boosting-type algorithms. Numerically, we carry out several simulations to show the promising performance of KReBooT in terms of its good generalization, near over-fitting resistance and structure constraints.

Keywords: Learning theory, kernel methods, boosting, re-scaling, truncation

1. Introduction

Kernel methods (Evgeniou et al., 2000), which map inputs of data to some kernel-based feature space to improve the learning performance of the classical linear approaches, have been widely used for regression in the last two decades. Excellent theoretical verifications in generalization performance of kernel methods (Cucker and Zhou, 2007; Steinwart and Christmann, 2008) are the most important driving force behind their popularity. In particular, optimal generalization error bounds for kernel ridge regression, kernel-based gradient descents, kernel-based spectral algorithms, and kernel-based conjugate gradient algorithms were established in (Caponnetto and De Vito, 2007; Steinwart et al., 2009), (Yao et al.,

*. Corresponding author

2007; Lin and Zhou, 2018a), (Bauer et al., 2007; Guo et al., 2017a), and (Blanchard and Krämer, 2016; Lin and Zhou, 2018b), respectively. To accommodate to the so-called "big data era", kernel methods were extended in the following three interesting ways:

• Scalability: it focuses on designing scalable variants of kernel methods to reduce the computational burden and storage requirements. Typical variants of kernel methods in this direction are the distributed kernel learning (Zhang et al., 2015), localized kernel learning (Meister and Steinwart, 2016) and kernel learning with sub-sampling (Williams and Seeger, 2000).

• Adaptivity: it aims to design adaptive and provable parameter-selection strategies to reduce the price of parameter-selection in kernel methods. Existing parameter-selection strategies for kernel methods are the cross-validation (Caponnetto and Yao, 2010), balancing principle (Lu et al., 2018), Lepskii principle Blanchard et al. (2019) and discrepancy principle (Celisse and Wahl, 2021) for kernel learning.

• Structure Constraints: it devotes to incorporating some structure information in kernel methods to reflect the a-priori knowledge of the learning task to either improve the generalization performance further or reduce the computation. Typical variants of kernel methods in this way include the kernel-based LASSO (Wang et al., 2007), kernel-based manifold regularization (Belkin et al., 2006) and kernel-based elastic net (Zou and Hastie, 2005).

In this paper, we mainly focus on the structure constraints issue of kernel methods in the framework of regression (Cucker and Zhou, 2007; Steinwart and Christmann, 2008). Our study is motivated by three interesting observations. At first, the representer theorem (Wahba, 1990) transforms the learning problems in usually infinite dimensional reproducing kernel Hilbert spaces (RKHSs) into m-dimensional linear regression problems, making the derived kernel-based estimator a linear combination of m computational units, where m is the size of training data. Therefore, given a query point, the testing complexity behaves linearly with m. In this way, the sparseness of a derived kernel-based estimator not only reflects the sparseness of regression functions (Shi et al., 2011), but also determines the testing complexity. Then, previous studies (Shi, 2013; Guo et al., 2017b) on sparseness-driven kernel methods including the kernel LASSO and kernel-based threshold algorithms did not encode the spareness structure in the derived estimator, making the derived generalization error bound not so tight and the sparsity of the derived estimator not so clear. Taking kernel LASSO for example, it usually requires large regularization parameter to enable the sparsity of the derived estimator (Lin et al., 2014), which may be questionable in generalization since too large regularization parameter may degrade the generalization performance due to the bias-variance trade-off principle (Cucker and Zhou, 2007). Finally, boosting (Freund, 1995) and its variants (Friedman, 2001; Zhang and Yu, 2005; Wang et al., 2019) have been widely used in numerous regression problems due to their sparseness-driven nature (Shalev-Shwartz et al., 2010). The problem is, however, that both the numerical convergence rate and the generalization error bound of the original boosting algorithm are not satisfactory. In particular, a generalization error bound of order $(\log m)^{-\mu}$ for some $\mu > 0$ was provided (Bickel et al., 2006) for the original boosting. Based on these three observations, our basic idea is to develop a novel variant of boosting to equip kernel methods to derive an estimator that possesses good generalization performance and controllable sparsity, simultaneously.

Regularization and *re-scaling* are two popular ideas to improve the performance of boosting. The idea of *Regularization* aims at controlling the step-size of boosting iterations and derives estimators with structure constraints (relatively small ℓ_1 norm). Regularized boosting via truncation (RTboosting) (Zhang and Yu, 2005) is a typical variant of boosting based on *regularization*. RTboosting controls the step-size in each boosting iteration via limiting the range of line search, which succeeds in improving the generalization performance of boosting via reducing the variance. However, to the best of our knowledge, fast numerical convergence rates were not provided for these variants. In particular, it can be found in (Zhang and Yu, 2005) that the numerical convergence rate of RTboosting is of an order $k^{-1/3}$, which is far worse than the optimal rate for nonlinear approximation $\mathcal{O}(k^{-1})$ (DeVore and Temlaykov, 1996). Here and hereafter, k denotes the number of boosting iterations. The idea of *re-scaling* focuses on multiplying a re-scaling parameter to the estimator of each boosting iteration to accelerate the numerical convergence rate of boosting and then to improve the generalization performance of boosting via reducing the bias. In particular, re-scaled boosting (Rboosting) (Wang et al., 2019) which shrinks the estimator obtained in the previous boosting iteration was shown to achieve the optimal numerical convergence rate under certain sparseness assumption. However, there aren't any guarantees for the structure (ℓ_1 norm) of the estimator, making the generalization error of derived estimator sensitive to the number of boosting iterations.

In this paper, we propose a novel kernel-based re-scaled boosting with truncation (KReBooT) to embody advantages of *regularization* and *re-scaling*, simultaneously. We find a close relation between the regularization parameter in RTboosting (Zhang and Yu, 2005) and re-scaling parameter in RBoosting (Wang et al., 2019) to accelerate the numerical convergence and control the ℓ_1 norm of the derived estimator. This together with the well developed integral operator technique in (Lin et al., 2017; Guo et al., 2017a) and a Talagrand's concentration inequality (Steinwart and Christmann, 2008) yields an almost optimal numerical convergence rate and a fast learning rate of KReBooT. In particular, the new algorithm can achieve a generalization error bound of order $\frac{1}{m}\log^4 m$ under some standard assumptions on the kernel and the data distribution. Due to the structure constraint, we also prove that the new algorithm is near overfitting resistant in the sense that the bias decreases inversely proportional to k, while the variance increases logarithmically with respect to k. We also conduct several numerical simulations to highlight the outperformance of KReBooT, compared with some widely used kernel methods. The numerical results are consistent with our theoretical claims and therefore verify our assertions.

The rest of paper is organized as follows. In the next section, we introduce detailed implementation and some basic properties of KReBooT. Section 3 provides convergence guarantees for KReBooT as well as its almost optimal numerical convergence rate. In Section 4, we derive fast learning rate for KReBooT in the framework of learning theory. Section 5 presents the numerical verifications for our theoretical assertions. In Section 6, we prove our main results.

2. Kernel-based Re-scaled Boosting with Truncation

Let $D = \{z_i\}_{i=1}^m = \{(x_i, y_i)\}_{i=1}^m$ be the set of samples with $x_i \in \mathcal{X}$ and $y_i \in \mathcal{Y}$, where $\mathcal{X}$ is a compact input space and $\mathcal{Y} \subseteq [-M, M]$ is the output space for some $M > 0$. Given a

bounded (strictly) positive definite kernel $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, denote by $\mathcal{H}_K$ the corresponding RKHS endowed with norm $\|\cdot\|_K$. The compactness of $\mathcal{X}$ and the boundedness of K imply $\kappa := \sqrt{\sup_{x\in\mathcal{X}} K(x,x)} < \infty$. Throughout this paper, we assume $\kappa \leq 1$ for the sake of brevity. Set $S := \{K_{x_i} : i = 1, \ldots, m\}$ with $K_x(\cdot) = K(\cdot, x)$. Let $\mathcal{H}_{K,D} := \{\sum_{i=1}^m a_i K_{x_i}(\cdot) : a_i \in \mathbb{R}\}$. For any vector $\mathbf{a} = (a_1, ..., a_m)^T$, write

$$\left\| \sum_{i=1}^m a_i K_{x_i}(\cdot) \right\|_{\ell_1} = \|\mathbf{a}\|_{\ell_1} = \sum_{i=1}^m |a_i|.$$

The well known representer theorem (Wahba, 1990) shows that all the aforementioned kernel-based algorithms build an estimator in $\mathcal{H}_{K,D}$. Thus, it is naturally to take $\mathcal{H}_{K,D}$ rather than $\mathcal{H}_K$ as the hypothesis space. It should be mentioned that the reason why we assume the (strictly) positive definiteness of K is to guarantee that the Gram matrix $\mathbb{K} = (K(x_i, x_j))_{i,j=1}^m$ is strictly positive definite for all randomly drawn $\{x_i\}_{i=1}^m$. Compared with classical kernel methods (Steinwart and Scovel, 2012) that only require bounded kernel and separable RKHS, there is an additional requirement on the strictly positive definiteness of the Gram matrix which can be guaranteed by the strictly positive definiteness of the kernel K in our analysis. In this way, S is a basis of $\mathcal{H}_{K,D}$ and consequently the dimesnsion of $\mathcal{H}_{K,D}$ is m, which implies that $\|\cdot\|_{\ell_1}$ is a norm on $\mathcal{H}_{K,D}$. Furthermore, for any function h and any convex set $\mathbb{A} \subset \mathcal{H}_{K,D}$, the (strictly) positive definiteness of K yields that there is a unique minimizer of the functional $f \mapsto \|h - f\|_m^2$ on $\mathbb{A}$. It is easy to check that the widely used kernels such as Gaussian kernels, Laplacian kernels and Sobolev kernels (Cucker and Zhou, 2007; Steinwart and Christmann, 2008) satisfy our conditions.

Kernel-based boosting aims at learning an estimator from $\mathcal{H}_{K,D}$ based on D. Combining kernel methods with boosting to yield kernel-based boosting algorithms is not a new idea in the realm of machine learning. It can data back to 2001, when Friedman (2001) formulated boosting as greedy approximation problems. As far as L_2-boosting is concerned, there are roughly three types of ways to combine boosting and kernel methods, which differ in building the sets of weaker learners. The first one, as (Friedman, 2001; Bühlmann and Yu, 2003) did, is to set the set of weak learners to be the family of some derived kernel estimators. In this way, L_2-boosting boils down to iteratively fitting the residual. In the recent paper (Lin et al., 2019), the authors combined L_2-boosting with kernel ridge regression to settle the well known saturation problem (Lo Gerfo et al., 2008) and present a near over-fitting resistant property for some specific kernels and data distributions. However, since it requires running kernel ridge regression several times, its computation complexities in both training and testing are relatively large. The second one, as (Yao et al., 2007; Raskutti et al., 2014) discussed, is to set $\mathcal{H}_{K,D}$ to be the set of weak learners. With this, L_2-boosting is similar as the classical kernel-based gradient descents (Yao et al., 2007; Wei et al., 2019). The main advantage of this method is its perfect generalization verifications in theory but the main drawback is due to its non-sparseness-driven property and difficulty on parameter selection, the later of which frequently requires delicate parameter selection strategy (Wei et al., 2019). The last one, as (Zhang and Yu, 2005; Shalev-Shwartz et al., 2010) declared, is to set $S = \{K_{x_i}(\cdot)\}_{i=1}^m$. This strategy coincides with the well known greedy algorithms (Barron et al., 2008) and dominates in reducing the computational burden and deducing sparse estimator (Zhang and Yu, 2005). However, the corresponding learning rates and numerical convergence rates are usually slow (Zhang and Yu, 2005).

In this paper, we focus on designing a kernel-based L_2-Boosting algorithm which is near over-fitting resistant, the generalization error increasing logarithmically with respect to the number of boosting iterations when over-fitting happens, and possesses excellent generalization performance by taking S as the set of weak learners. Our basic idea is to combine the classical truncation operator in (Zhang and Yu, 2005) to reduce the variance and a recently developed re-scaling technique in (Wang et al., 2019) to accelerate the numerical convergence rate and then reduce the bias. We thus nominate the new algorithm as kernel-based re-scaled boosting with truncation (KReBooT). Given a set of re-scaling parameters $\{\alpha_k\}_{k=1}^{\infty}$ with $\alpha_k \in (0,1)$ and a set of non-decreasing step sizes $\{l_k\}_{k=1}^{\infty}$, KReBooT starts with $f_{D,0} = 0$ and then iteratively implements the following two steps for $k \geq 1$:
Step 1 (Projection of gradient): Find a $g_k^* \in S$ such that

$$g_k^* := g_{k,D}^* := \arg\max_{g \in S} |\langle y - f_{D,k-1}, g \rangle_m|, \tag{1}$$

where $\langle f, g \rangle_m := \frac{1}{m} \sum_{i=1}^m f(x_i) g(x_i)$ and we abuse the notation a little bit to write y as a function satisfying $y(x_i) = y_i$.
Step 2 (Line search with re-scaling and truncation): Define

$$f_{D,k} := (1 - \alpha_k) f_{D,k-1} + \beta_k^* g_k^*, \tag{2}$$

where

$$\beta_k^* := \arg\min_{\beta \in \Lambda_k} \|(1 - \alpha_k) f_{D,k-1} + \beta g_k^* - y\|_m^2, \tag{3}$$

$\Lambda_k := [-\alpha_k l_k, \alpha_k l_k]$ and $\|f\|_m = \sqrt{\langle f, f \rangle_m}$.

Compared with the classical boosting algorithm in which the line search is conducted on $\mathbb{R}$ rather than Λ_k and $\alpha_k = 0$, KReBooT involves two crucial operators, i.e., re-scaling and truncation, to control the structure of the derived estimator. In particular, we can derive the following structure constraint for KReBooT.

Lemma 1 *Let $f_{D,k}$ be defined by (2). If $\{l_k\}_{k=1}^{\infty}$ with $l_k \geq 0$ are nondecreasing and $0 < \alpha_k < 1$, then*

$$\|f_{D,k}\|_{\ell_1} \leq l_k, \qquad \forall k = 0, 1, \ldots.$$

Proof It follows from $\Lambda_1 = [-\alpha_1 l_1, \alpha_1 l_1]$, $\alpha_1 \leq 1$ and $f_{D,0} = 0$ that

$$\|f_{D,1}\|_{\ell_1} \leq (1 - \alpha_1) \|f_{D,0}\|_{\ell_1} + l_1 \leq l_1.$$

If we assume $\|f_{D,k}\|_{\ell_1} \leq l_k$, then (2) yields

$$\begin{aligned} \|f_{D,k+1}\|_{\ell_1} &\leq (1 - \alpha_{k+1}) \|f_{D,k}\|_{\ell_1} + \alpha_{k+1} l_{k+1} \\ &\leq (1 - \alpha_{k+1}) l_{k+1} + \alpha_{k+1} l_{k+1} = l_{k+1}. \end{aligned}$$

This proves Lemma 1 by induction. ■

Lemma 1 shows that the ℓ_1 norm of the KReBooT estimator can be bounded by the step-size parameter l_k. At the first glance, a small ℓ_1 norm does not necessary implies the

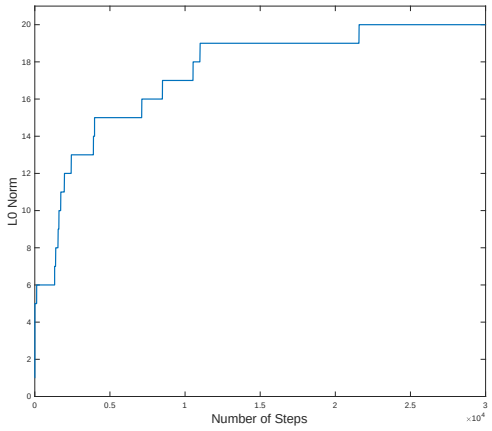

Figure 1: The change of support with respect to the boosting iteration

sparsity of the derived KReBooT estimator since small ℓ_1 norm can be fulfilled allowing the full set S to enter the final estimator with small coefficients. However, noticing that the ℓ_1 norm is a better approximation of the ℓ_0 (semi-)norm than the classical RKHS norm, small ℓ_1 norm possibly results in small ℓ_0 norm that is beyond the capability of the widely used RKHS norm in classical kernel methods (Guo et al., 2017a). Throughout the paper, the ℓ_0 (semi-)norm denotes the Cardinality of the support set of $f_{D,k}$. Moreover, due to the algorithmic design of KReBooT, different iterations are permitted to choose the same atom in S, which implies that the ℓ_0 norm of the derived estimator may be much smaller than k. Figure 1, whose detailed simulation setting can be found in Section 5, shows the change of support when boosting iteration happens. It can be found in the figure that the ℓ_0 norm of $f_{D,k}$ is smaller than 20 even when k approaches 30000. It would be interesting to theoretically study the change of support with respect to k, just as (Ehrlinger and Ishwaran, 2012) did for L_2-boosting with the traditional features. We will leave it for a future study since this paper pays more attention to the numerical convergence rate and generalization capability of KReBooT.

From the above descriptions, there are totally three types of parameters in KReBooT: re-scaling parameter α_k, step-size parameter l_k and iteration number k, in which l_k is imposed to control the sparsity and α_k is adopted to accelerate the numerical convergence rate. We will present detailed parameter-selection strategies after the theoretical analysis and show that the difficulty of selecting each parameter is much less than other variants of boosting (Zhang and Yu, 2005; Xu et al., 2017). For arbitrary $g \in S$ and $\beta \in \Lambda_k$, we have

$$\|[(1-\alpha_k)f_{D,k-1} + \beta g] - y\|_m^2 = \frac{1}{m}\sum_{i=1}^m [(1-\alpha_k)f_{D,k-1}(x_i) - y_i]^2 + \frac{\beta^2}{m}\sum_{i=1}^m g^2(x_i)$$

$$- \quad \frac{2\beta}{m}\sum_{i=1}^m [(1-\alpha_k)f_{D,k-1}(x_i) - y_i]g(x_i).$$

Direct computation then yields

$$\beta_k^* := sign(\langle r_{k-1}, g_k^* \rangle_m) \min\left\{ \frac{|\langle r_{k-1}, g_k^* \rangle_m|}{\|g_k^*\|_m^2}, \alpha_k l_k \right\}, \tag{4}$$

Algorithm 1 KReBooT

Input: $D = \{(x_i, y_i)\}_{i=1}^m$, kernel $K(\cdot, \cdot)$.
Parameters: Re-scaling parameter $\alpha_k \in (0, 1)$, step size $l_k \in \mathbb{R}_+$, and number of iterations $k = 1, 2, \ldots,$

for $k = 1, \ldots$ **do**
▶ (Projection of Gradient) Find $g_k^* \in S$ satisfying (1).
▶(Line search with re-scaling and truncation) Define

$$f_{D,k} := (1 - \alpha_k) f_{D,k-1} + \beta_k^* g_k^*,$$

where β_k^* is obtained by (4) and $f_{D,0} = 0$.
end for

where $r_{k-1} := r_{D,k-1} := y - (1 - \alpha_k) f_{D,k-1}$ and $sign(\cdot)$ is the sign function. With these, we summary the detailed implementation of KReBooT in Algorithm 1.

We conclude this section by discussing the computational cost and storage requirement for KReBooT. In the Projection of gradient step, it requires to compute $\langle y, k_{x_i} \rangle_m$ and $\langle K_{x_j}, K_{x_i} \rangle_m$ for any $1 \le i, j \le m$ and thus needs $\mathcal{O}(m^2)$ float computations. If we store the above $m^2 + m$ quantities, then it totally requires $\mathcal{O}(m^2 + km)$ computational complexity in the Projection of gradient step when KReBooT iterates k-times, since

$$\langle y - f_{D,k-1}, g \rangle_m = \langle y, g \rangle - \sum_{i=1}^m a_i \langle K_{x_i}, g \rangle_m$$

for $f_{D,k_1} = \sum_{i=1}^m a_i K_{x_i}$ and $g = K_{x_j}$ for $j = 1, \ldots, m$. In the line search step, it is obvious $\mathcal{O}(km)$ float computations are required. Therefore, it totally needs $\mathcal{O}(m^2 + km)$ float computations for running KReBooT with k steps.

3. Numerical Convergence of KReBooT

In this section, we conduct the numerical convergence analysis for KReBooT. At first, we present a sufficient condition for $\{\alpha_k\}_{k=1}^\infty$ to guarantee the convergence of Algorithm 1 in the following theorem.

Theorem 2 *Assume $|y_i| \le M$ and $\kappa \le 1$. Let $\{f_{D,k}\}_{k=0}^\infty$ be defined in Algorithm 1 with nondecreasing sequence $\{l_k\}_{k=1}^\infty \subset (0, \infty)$ and nonincreasing sequence $\{\alpha_k\}_{k=1}^\infty \subset (0, 1)$. If*

$$\lim_{k \to \infty} \alpha_k = 0, \quad \text{and} \quad \sum_{k=1}^\infty \alpha_k = \infty, \tag{5}$$

then,

$$\lim_{k \to \infty} f_{D,k} = \begin{cases} \arg \min\limits_{f \in \mathcal{H}_{K,D}} \|f - y\|_m^2, & \text{when } l_k \uparrow \infty, \\ \arg \min\limits_{f \in B_L} \|f - y\|_m^2, & \text{when } l_k \equiv L, \end{cases}$$

where $B_L = \{f \in \mathcal{H}_{K,D} : \|f\|_{\ell_1} \le L\}$.

Since the range of line search of KReBooT is $[-\alpha_k l_k, \alpha_k l_k]$, $\alpha_k \to 0$ together with l_k satisfying $\lim_{k\to\infty} \alpha_k l_k = 0$ implies the convergence of $f_{D,k}$. An extreme case is to set $\alpha_k \equiv 0$, which certainly guarantees the convergence but does not guarantee the effectiveness of the boosting iteration. Differently, the condition $\sum_{k=1}^{\infty} \alpha_k = \infty$ guarantees the effectiveness of the iteration and controls where the algorithm converges. For different type of l_k, KReBooT converges either to an empirical kernel-based least-squares solution or a kernel-based LASSO solution.

Due to the special iteration rule of KReBooT, its numerical convergence rate depends heavily on the re-scaling parameter α_k. If α_k is too large, then the re-scaling operator offsets the effectiveness of the previous boosting iteration, making the numerical convergence rate slow. On the contrary, if α_k is too small, then step sizes of the line search are also very small, reducing the effectiveness of boosting iterations. Therefore, a suitable selection of the re-scaling parameter α_k is highly desired in KReBooT. In the following theorem, we show that KReBooT with $\alpha_k = \frac{2}{k+2}$ can achieve the optimal numerical convergence rate of nonlinear approximation.

Theorem 3 *Assume $|y_i| \le M$ and $\kappa \le 1$. If $h \in \mathcal{H}_{K,D}$ satisfying $\|h\|_{\ell_1} < \infty$, $\alpha_k = \frac{2}{k+2}$ and $\{l_k\}_{k=1}^{\infty}$ is a sequence of nondecreasing positive numbers satisfying $\lim_{k\to\infty} l_k = \infty$, then for any $k \ge 1$,*

$$\|y - f_{D,k}\|_m^2 - \|y - h\|_m^2 \le \left[M^2(k_h^* + 2)^2 + 4(M + \|h\|_{\ell_1})^2\right]/(k+2), \tag{6}$$

where k_h^* is the smallest positive integer satisfying $l_{k_h^*} \ge \|h\|_{\ell_1}$.

In (6), the convergence rate depends on k_h^* and $l_{k_h^*}$. For a given h with $\|h\|_{\ell_1} < \infty$ and a sequence of nondecreasing positive numbers $\{l_k\}_{k=1}^{\infty}$ with $\lim_{k\to\infty} l_k = \infty$, there always exists a constant k_h^* such that $l_{k_h^*} \ge \|h\|_{\ell_1}$. Under this circumstance, k_h^* and $l_{k_h^*}$ can be regarded as constants in the estimate. However, it should be mentioned that for k satisfying $l_k < \|h\|_{\ell_1}$, the boosting iteration in Algorithm 1 is not effective and the algorithm requires increasing property of $\{l_k\}_{k=1}^{\infty}$. Once $l_k \ge \|h\|_{\ell_1}$ for some k, then the algorithm converges of an order k^{-1}. In other words, there are actually two regimes for KReBooT approximation. The first one focuses on small k where the output estimator is biased, while the second one starts when k is larger than k^* and the bound (6) starts to be meaningful. Therefore, the selection of ℓ_k determines the boundary of these two regimes. Theorem 3 devotes to an arbitrary $h \in \mathcal{H}_{K,D}$ satisfying $\|h\|_{\ell_1} < \infty$, implying the following corollary directly.

Corollary 4 *Assume $|y_i| \le M$ and $\kappa \le 1$. If*

$$h^* = \arg \min_{\|f\|_{\ell_1} \le \bar{c}_0} \arg \min_{f \in \mathcal{H}_{K,D}} \frac{1}{m} \sum_{i=1}^{m} (f(x_i) - y_i)^2$$

for some $\bar{c}_0 > 0$, then

$$\|y - f_{D,k}\|_m^2 - \|y - h^*\|_m^2 \le \left[M^2(k_{h^*}^* + 2)^2 + 4(M + \bar{c}_0)^2\right]/(k+2).$$

The above corollary compares the numerical convergence performance of KeReBooT and the kernel Lasso and shows that only a gap of order k^{-1} in this two strategies. If the sparsity information of h, $\|h\|_{\ell_1}$, is known, then we can set $l_k = \|h\|_{\ell_1}$ directly for all $k = 1, 2, \ldots$. In this way, we have $k_h^* = 1$ and $l_{k_h^*} = \|h\|_{\ell_1}$, which yields the following corollary.

Corollary 5 *Assume $|y_i| \leq M$ and $\kappa \leq 1$. If $h \in \mathcal{H}_{K,D}$ satisfying $\|h\|_{\ell_1} < \infty$, $\alpha_k = \frac{2}{k+2}$ and $l_k = \|h\|_{\ell_1}$, then there holds $\|f_{D,k}\|_{\ell_1} \leq \|h\|_{\ell_1}$ and*

$$\|y - f_{D,k}\|_m^2 - \|y - h\|_m^2 \leq \left[(9M^2 + 4(M + \|h\|_{\ell_1})^2\right] k^{-1}. \tag{7}$$

If $\|h\|_{\ell_1}$ is unknown before the learning process, we recommend to set $l_k = c_0 \log(k+1)$ for some $c_0 > 0$ and get the following lemma, to balance the numerical convergence rate and ℓ_1 norm of the derived estimator.

Corollary 6 *Assume $|y_i| \leq M$ and $\kappa \leq 1$. If $h \in \mathcal{H}_{K,D}$ satisfying $\|h\|_{\ell_1} < \infty$, $\alpha_k = \frac{2}{k+2}$ and $l_k = c_0 \log(k+1)$, then $\|f_{D,k}\|_{\ell_1} \leq c_0 \log(k+1)$ and*

$$\|y - f_{D,k}\|_m^2 - \|y - h\|_m^2 \leq \left[M^2(e^{\|h\|_{\ell_1}/c_0} + 2)^2 + 4(M + \|h\|_{\ell_1})^2\right] k^{-1}, \tag{8}$$

where c is a constant depending only on M and $\|h\|_{\ell_1}$.

Recalling that the optimal rate of nonlinear approximation under the same conditions as Theorem 3 is $\mathcal{O}(k)$, the above two corollaries show that KReBooT succeeds in achieving this optimal rate and is essentially better the original Boosting whose numerical convergence rate lies in $(k^{-0.3796}, k^{-0.364})$ and RTBoosting whose numerical convergence rate is $\mathcal{O}(k^{-1/3})$ (Zhang and Yu, 2005). Though RBoosting (Bagirov et al., 2010; Xu et al., 2017) can also reach this optimal convergence rate $\mathcal{O}(k^{-1})$, there lacks a tight bound of the ℓ_1 norm, making it sensitive to the re-scaling parameter and number of iterations.

4. Generalization Error Analysis

In this section, we are interested in deriving tight generalization error bounds for KReBooT. Our analysis is carried out in the framework of statistical learning theory (Cucker and Zhou, 2007), where $D = \{z_i\}_{i=1}^m = \{(x_i, y_i)\}_{i=1}^m \subset \mathcal{Z}$ are assumed to be drawn independently and identically according to an unknown joint distribution $\rho := \rho(x, y) = \rho_X(x)\rho(y|x)$ with ρ_X the marginal distribution and $\rho(y|x)$ the conditional distribution. The learning performance of an estimator f is measured by the generalization error $\mathcal{E}(f) := \int_{\mathcal{Z}}(f(x) - y)^2 d\rho$. Noting that the regression function defined by $f_\rho(x) = E[y|X = x]$ minimizes the generalization error, our target is then to learn a function f_D to approximate f_ρ such that

$$\mathcal{E}(f_D) - \mathcal{E}(f_\rho) = \|f_D - f_\rho\|_\rho^2 \tag{9}$$

is as small as possible, where $\|\cdot\|_\rho$ is the norm of the space $L^2_{\rho_X}$.

To quantify the learning performance of KReBooT, some a-priori information concerning the kernel K, the marginal distribution ρ_X and regression function f_ρ should be given. Our first assumption describes the smoothness of the kernel K and the regularity of the marginal

distribution ρ_X via introducing the following integral operator. Define $L_K : L^2_{\rho_X} \to L^2_{\rho_X}$ (or $\mathcal{H}_K \to \mathcal{H}_K$) as

$$L_K f := \int_{\mathcal{X}} K_x f(x) d\rho_X. \tag{10}$$

Since K is (strictly) positive definite, L_K is a positive operator. Let $\{(\mu_\ell, \phi_\ell)\}_{\ell=1}^{\infty}$ be a set of normalized eigenpairs of L_K with $\{\mu_\ell\}_{\ell=1}^{\infty}$ arranging in a non-increasing order. The following assumption quantifies the decay rate of μ_ℓ.

Assumption 1 *For $0 < s < 1$ and some $c > 0$, we assume*

$$\mu_\ell \le c\ell^{-1/s}, \qquad \forall \ell \ge 1. \tag{11}$$

Since the estimator derived by Algorithm 1 is always in $\mathcal{H}_{K,D}$ with controllable ℓ_1 norm, its generalization performance depends heavily on the kernel and consequently s in Assumption 1. If ρ_X is the uniform distribution on the unit interval $\mathbb{I}$ and K is the reproducing kernel of the Sobolev space $W_\alpha(\mathbb{I})$ with $\alpha > 1/2$, i.e., the Sobolev spline kernel $S_\tau(x, x') = \frac{2\pi}{\Gamma(\alpha)} \mathbb{B}_{\alpha-1/2}(|x - x'|)(|x - x'|/2)^{\alpha-1/2}$, then (11) always holds with $s = u^{-1/(2\alpha)}$, where $\mathbb{B}_v$ is the modified Bessel function of the second type of order v. More examples for ρ_X and K satisfying Assumption 1 with specific s can be found in (Cucker and Zhou, 2007; Steinwart and Christmann, 2008). Assumption 1 is widely used in bounding generalization error bounds for numerous kernel-based algorithms (Caponnetto and De Vito, 2007; Steinwart et al., 2009; Raskutti et al., 2014; Zhang et al., 2015; Lin et al., 2017; Lu et al., 2018; Lin et al., 2019).

Our second assumption devotes to the sparseness of the regression function f_ρ with respect to the set of weak learners S.

Assumption 2 *Let $r \ge 0$. For any $0 \le \tau < 1$, with confidence $1 - \tau$, there exists an $h_{D,\rho} \in \mathcal{H}_{K,D}$ satisfying*

$$\|f_\rho - h_{D,\rho}\|_\rho \le c_1(\tau) m^{-r}, \qquad \text{and } \|h_{D,\rho}\|_{\ell_1} \le c_2(\tau), \tag{12}$$

where $c_1(\tau), c_2(\tau) \ge 1$ are two constants depending only on τ.

Given the set of weak learners S, a standard a-priori information on the sparseness of the regression in the boosting literature (Zhang and Yu, 2005; Bickel et al., 2006) is

$$f_\rho \in \operatorname{span}(S) \qquad \text{and} \quad \|f_\rho\|_{\ell_1} \le \tilde{C}_1 \tag{13}$$

for some constant $\tilde{C}_1 > 0$, showing that f_ρ can be represented by a linear combination of weak learners with small ℓ_1 norm. Assumption 2 is different from (13) in two aspects. On one hand, the hypothesis space $\mathcal{H}_{K,D}$ in this paper depends on the data set D and thus is random, making our analysis be carried out in probability arguments. On the other hand, Assumption 2 extends (13) via considering $f_\rho \notin \mathcal{H}_{K,D}$ by presenting an approximation rate of order m^{-r} in (12). In this way, (12) reflects the sparseness of the regression function with respect to $\mathcal{H}_{K,D}$. Noting further $\|f\|_\rho \le \|f\|_\infty$, Assumption 2 is also looser than the

widely used sparseness assumption in (Barron et al., 2008; Bagirov et al., 2010; Xu et al., 2017; Wang et al., 2019): there exists an $h_\rho \in \text{span}(S)$ such that

$$\|f_\rho - h_\rho\|_\infty \le \tilde{C}_2 m^{-r} \qquad \text{and} \quad \|h_\rho\|_{\ell_1} \le \tilde{C}_3 \tag{14}$$

for some constant $\tilde{C}_2, \tilde{C}_3 \ge 0$. The following proposition shows that the sparseness assumption always holds under some regularity condition of f_ρ.

Proposition 7 *If Assumption 1 holds with some $c > 0$ and $0 < s < 1$ and there exists an $h_\rho \in L^2_{\rho_X}$ such that $f_\rho = L_K h_\rho$, then Assumption 2 holds with $r = \frac{1}{2(s+1)}$, $c_1(\tau) = \tilde{c}_1 \|h_\rho\|_\rho \log \frac{4}{\tau}$ and $c_2(\tau) = 4\|h_\rho\|_\rho \log \frac{2}{\tau}$ where $\tilde{c}_1 = 3 + 2\sqrt{c + \frac{c}{1-s} c^{1-1/s}}$.*

A standard assumption for kernel methods is that $f_\rho \in \mathcal{H}_K$ (Zhang et al., 2015), which is equivalent to the condition that there exists an $h_\rho \in L^2_{\rho_X}$ such that $f_\rho = L_K^{1/2} h_\rho$ (Caponnetto and De Vito, 2007). The regularity condition in Proposition 7 is stricter than this standard assumption. For instance, if K is a Sobolev kernel associated with the Sobolev space $W_\alpha(\mathbb{I})$ satisfying $\sigma_k \sim k^{-2\alpha}$, then $f_\rho \in \mathcal{H}_K$ is equivalent to $f_\rho \in W_\alpha(\mathbb{I})$ while $f_\rho = L_K h_\rho$ implies $f_\rho \in W_{2\alpha}(\mathbb{I})$. Since $L_K h_\rho = \int_{\mathcal{X}} K_x h_\rho(x) d\rho_X$, the regularity condition $f_\rho = L_K h_\rho$ can be regarded as a population version of the assumption $f_\rho \in \mathcal{H}_{K,D}$. Besides the assumption $f_\rho = L_K h_\rho$ for some $h_\rho \in L^2_{\rho_X}$ in Proposition 7 requiring f_ρ to be in a subset of $\mathcal{H}_K$, there are also some $f_\rho \notin \mathcal{H}_K$ satisfying Assumption 2. For example, it is easy to derive that $f_\rho = L_K h_\rho + g(x)$ with $g \notin \mathcal{H}_K$ and $\|g\|_\infty \le m^{-r}$ satisfies Assumption 2. It would be interesting and useful to provide more realistic conditions on f_ρ to satisfy Assumption 2.

By the help of the above two assumptions, we present our third main result in the following theorem to quantify the learning performance of KReBooT.

Theorem 8 *Let $0 < \delta < 1$, $|y_i| \le M$, $\kappa \le 1$ and $f_{D,k}$ be defined by Algorithm 1 with $\alpha_k = \frac{2}{k+2}$ and $l_k = c_0 \log(k+1)$ for some $c_0 > 0$. If Assumption 2 and Assumption 1 hold with some $0 < s < 1$ and $c > 0$, then for any $k \ge \min\left\{m^{\frac{1}{1+s}}, m^{(1-s)r+\frac{1}{2}}\right\}$, with confidence $1-\delta$, there holds*

$$\mathcal{E}(f_{D,k}) - \mathcal{E}(f_\rho) \le C(\delta) \left(\log(k+1)\right)^4 \max\left\{m^{-\frac{1}{1+s}}, m^{-(1-s)r-\frac{1}{2}}\right\} \log^2 \frac{6}{\delta}, \tag{15}$$

where

$$C(\delta) := C'(c_1(\delta/6) + k^*_{\delta/6}(k^*_{\delta/6} + l^2_{k^*_{\delta/6}}) + (c_2(\delta/6))^2 (c_1(\delta/6))^{(1-s)/2})^2, \tag{16}$$

k^*_τ is the smallest positive integer satisfying $l_{k^*_\tau} \ge c_2(\tau)$, and C' is a constant depending only on M, s and c.

Theorem 8 shows that under Assumption 2 with $r \ge \frac{1}{2(1+s)}$ and Assumption 1 with a sufficiently small s, KReBooT achieves a generalization error bound of order $m^{-1}(\log k)^4$ with high probability, provided $k \ge \min\left\{m^{\frac{1}{1+s}}, m^{(1-s)r+\frac{1}{2}}\right\}$. It should be mentioned that it is a new record for boosting-type algorithms. In particular, under the slightly stricter condition (14), the generalization error of RTboosting (Zhang and Yu, 2005) is worse than $\mathcal{O}(m^{-1/2})$ while that of Rboosting (Barron et al., 2008) is $\mathcal{O}(m^{-1/2})$. RBoosting can reach

Table 1: Parameter-selection among different boosting algorithms

Algorithm	KReBooT				
Parameters	Parameter k	Parameter l_k	Parameter α_k	Convergence rate	Generalization error
Selection	$C_0 m$	$c_0 \log(k+1)$	$\frac{2}{k+2}$	$\mathcal{O}(k^{-1})$	$\mathcal{O}(m^{-1/(1+s)} \log^4 m)$
Algorithm	RTBoosting				
Parameter	Parameter k	Parameter l_k	Null	Convergence rate	Generalization error
Selection	$C\sqrt{m}$	$ck^{-2/3}$	Null	$\mathcal{O}(k^{-1/3})$	$\mathcal{O}(m^{-1/3})$
Algorithm	RBoosting				
Parameter	Parameter k	Null	Parameter α_k	Convergence rate	Generalization error
Selection	$C\sqrt{m}$	Null	$\frac{c}{c+k}$	$\mathcal{O}(k^{-1})$	$\mathcal{O}(m^{-1/2})$
Algorithm	Boosting				
Parameter	Parameter k	Null	Null	Convergence rate	Generalization error
Selection	$C(\log m)^{6\mu}$	Null	Null	$\mathcal{O}(k^{-1/3})$	$\mathcal{O}((\log m)^{-\mu})$

a similar generalization error as Theorem 8 (Wang et al., 2019) but requires additional compactness assumption on the hypothesis space, which is usually difficult to be satisfied. The main reason for the breakthrough of KReBooT is due to the structure constrains presented in Lemma 1, showing that the ℓ_1 norm is controllable when the boosting iterations happen, and the optimal numerical convergence rate established in Theorem 3, demonstrating that the number of boosting iterations in KReBooT to achieve a specific accuracy is almost the smallest. To guarantee the good learning performance of KReBooT, there are two requirements on the number of iterations, i.e., $k \geq \min\left\{m^{\frac{1}{1+s}}, m^{(1-s)r+\frac{1}{2}}\right\}$ and k is not exponential with respect to m. In this way, a selection of $k = C_0 m$ is recommended. From (15), there is an additional logarithmic term $\log^4(k+1)$ in the bound of generalization error. This is due to $\|f_{D,k}\|_{\ell_1} \leq l_k = c_0 \log(k+1)$ according to Lemma 1, which results in an additional logarithmic term with respect to k in bounding the sample error (variance). If $c_2(\tau)$ in Assumption 2 is known before the learning process, then we can set $l_k \equiv c_2(\tau)$ to remove such a logarithmic term.

Noting that there are three tunable parameters in Algorithm 1, that is, α_k, l_k and k. Our theoretical analysis and experimental verification below show that KReBooT is stable with respect to α_k that can be fixed to be $\frac{2}{k+2}$ before the learning process. Moreover, we can set $l_k = c_0 \log(k+1)$ to guarantee the good structure of the KReBooT estimator. Here, c_0 is a parameter which affects the constant $C(\delta)$ in (15). In particular, the selection of c_0 is somewhat important but not difficult in the sense that it is easy to get an appropriate c_0 by using some standard parameter-selection strategies such as "hold-out" (Caponnetto and Yao, 2010) or cross-validation (Györfy et al., 2002). The selection of k depends on c_0. If c_0 is extremely large, ∞ for example, then the truncation operator does not make sense and the performance of algorithm is sensitive to k. If c_0 is suitable, it follows from Theorem 8 that a large k, comparable with m or larger, is good enough. Thus, in the practical implementation of KReBooT, we suggest to set $\alpha_k = \frac{2}{k+2}$, k to be large and $l_k = c_0 \log(k+1)$ and $k = C_0 m$ with c_0, C_0 being tunable parameters. Under this circumstance, there are two parameters c_0, C_0 in the new algorithm, which is similar as other variants of regularized boosting algorithms such as RTboosting and Rboosting. Recalling Theorem 8, the selection of C_0 does not significantly affects the generalization error of KReBooT. Table 1 presents the detailed parameter selection strategies and corresponding numerical convergence rates and

Table 2: Comparison among kernel methods under assumptions of Corollary 9, in which Relation means that the relation between generalization error and parameter when over-fitting happens

Algorithm	Generalization	ℓ_1 norm	Relation	Computation
KReBoot	$m^{-1/(1+s)}\log m$	$\log m$	Logarithmic	m^2
KRR	$m^{-2/(2+s)}$	Null	Algebraic	m^3
KGD	$m^{-2/(2+s)}$	Null	Algebraic	km^2
KCG	$m^{-2/(2+s)}$	Null	Algebraic	m^3
KPCA	$m^{-2/(2+s)}$	Null	Algebraic	m^3
KLASSO	$m^{-1/(2+2s)}$	$\log m$	Algebraic	m^3

generalization error under Assumption 2 with $r \geq \frac{1}{2(s+1)}$ and Assumption 1 for KReBooT, RTboosting (Zhang and Yu, 2005), RBoosting (Wang et al., 2019) and Boosting (Bickel et al., 2006).

Theorem 8 shows that under the sparseness assumption (12) and capacity assumption (11), KReBooT possesses the excellent generalization performance and is near over-fitting resistant with respect to the number of boosting iterations. The following corollary can be derived from Proposition 7 and Theorem 8 yields the following corollary.

Corollary 9 *Let $0 < \delta < 1$ $|y_i| \leq M$, $\kappa \leq 1$, and $f_{D,k}$ be defined by (2) with $\alpha_k = \frac{2}{k+2}$ and $l_k = c_0 \log(k+1)$ for some $c_0 > 0$. If Assumption 1 holds with some $c > 0$ and $0 < s < 1$, there exists an $h_\rho \in L^2_{\rho_X}$ such that $f_\rho = L_K h_\rho$, and $k = C_0 m$, then for $k = C_0 m$, with confidence $1-\delta$, there holds*

$$\mathcal{E}(f_{D,k}) - \mathcal{E}(f_\rho) \leq C\,(\log m)^4\, m^{-\frac{1}{1+s}} \log^5 \frac{24}{\delta}, \tag{17}$$

where C is a constant independent of m, k or δ.

Under the same assumptions, our derived generalization error in (17) is much better than that of kernel-based LASSO (KLASSO) (Shi et al., 2011; Shi, 2013; Guo and Shi, 2013), where generalization errors of order $m^{-1/2}$ were provided. It should be mentioned that the derived bound in Corollary 9 is worse than some existing kernel approaches like kernel ridge regression (KRR) (Caponnetto and De Vito, 2007; Steinwart et al., 2009), kernel gradient descent (KGD) (Lin and Zhou, 2018a), kernel conjugate descent (KCG) (Blanchard and Krämer, 2016) and kernel PCA (KPCA) (Guo et al., 2017a), in which a learning rate of order $m^{-\frac{2}{2+s}}$ is derived. Table 2 presents a detailed comparison among KReBooT, KRR, KGD, KLASSO, KCG and KPCA.

Compared with the methods prsesented in Table 2, it can be found that KReBooT possesses three advantages: low computation, over-fitting resistance and large applicable range by noting Proposition 7. The relatively bad generalization error of KReBooT is due to the structure constraints on the derived estimator as in Lemma 1, i.e., the essential hypothesis space of KReBooT is $B_{K,c_0\log(k+1)} = \{f = \sum_{i=1}^m a_i K_{x_i} : \sum_{i=1}^m |a_i| \leq c_0 \log k\}$ while that of KRR, KGD, KCG and KPCA is $\{f = \sum_{i=1}^m a_i K_{x_i} : \|f\|_K \leq m^a\}$ for some

$a \geq 0$. This generalization error bound does not implies that KReBooT always performs worse than the mentioned kernel methods since the conclusion in Corollary 9 is made in the worst case for f_ρ. In the following corollary derived from Theorem 8 directly, we show that KReBooT performs at least essentially worse than others for other types of regression functions.

Corollary 10 *Let $0 < \delta < 1$, $|y_i| \leq M$, $\kappa \leq 1$ and $f_{D,k}$ be defined by Algorithm 1 with $\alpha_k = \frac{2}{k+2}$, $l_k = c_0 \log(k+1)$ and $k = C_0 m$ for some $c_0, C_0 > 0$. If Assumption 1 holds with some $0 < s < 1$ and $c > 0$ and $f_\rho \in \mathcal{H}_{K,D}$ satisfying $\|f_\rho\|_{\ell_1} < \infty$, then with confidence $1-\delta$, there holds*

$$\mathcal{E}(f_{D,k}) - \mathcal{E}(f_\rho) \leq C' m^{-\frac{1}{1+s}} \log^4 m \log^2 \frac{6}{\delta}, \tag{18}$$

where C' is a constant depending only on $\|f_\rho\|_{\ell_1}$, c_0 and C_0.

Since it is difficult for classical kernel methods to derive satisfactory generalization errors under the sparseness assumption (Cucker and Zhou, 2007), they only regard $f_\rho \in \mathcal{H}_{K,D}$ and $\|f_\rho\|_{\ell_1} < \infty$ to be $f_\rho \in \mathcal{H}_K$. In this way, the best generalization error bounds of them are of order $\mathcal{O}(m^{-\frac{1}{1+s}})$ (Lin et al., 2017). It should be mentioned that Nyström regularization (Rudi et al., 2015; Lu et al., 2019) is also a preferable strategy to reduce the computational complexity without sacrificing the generalization performance of kernel methods. The problem is, however, that there is an additional parameter, the sub-sampling ratio, involved in Nyström regularization. Though the sub-sampling ratios are quite small in practice, theoretical assessments derived in (Rudi et al., 2015) showed that it should be larger than $m^{\frac{1}{1+s}}$ to achieve the optimal generalization error bounds, which would be near to 1 for some kernel and ρ_X satisfying Assumption 1 with s near to 0.

5. Experiments

5.1 Simulated data experiments

In this subsection, we conduct several simulations to show the performance of KReBooT. In the following Simulations, we consider the regression model as

$$y_i = g(x_i) + \varepsilon_i, \quad i = 1, 2, \cdots, N, \tag{19}$$

where $\{x_i\}_{i=1}^m$ are i.i.d. drawn according to the uniform distribution in the interval $(0, 1)$, $\{\varepsilon_i\}$ are the i.i.d. random Gaussian variables, and

$$g(x) := h_2(\|x\|_2) := \begin{cases} (1 - \|x\|_2)^6 (35\|x\|_2^2 + 18\|x\|_2 + 3), & 0 < \|x\|_2 \leq 1, x \in \mathbb{R}^3, \\ 0, & \|x\|_2 > 1. \end{cases}$$

The kernel used for the proposed KReBooT is chosen as $K(x, x') = h_3(\|x - x'\|_2)$ with

$$h_3(\|x\|_2) := \begin{cases} (1 - \|x\|_2)^4 (4\|x\|_2^2 + 1), & 0 < \|x\|_2 \leq 1, x \in \mathbb{R}^3, \\ 0, & \|x\|_2 > 1. \end{cases}$$

The reason why we make such choices of $g(\cdot)$ and $K(\cdot)$ is to guarantee the assumptions of Corollary 9 (Chang et al., 2017).

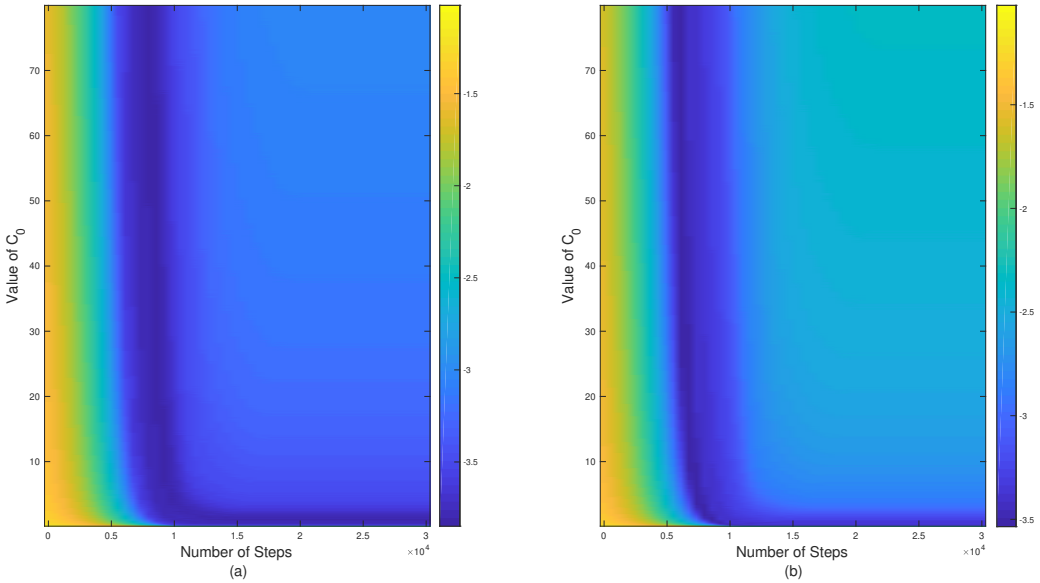

Figure 2: The visualization of testing MSE of the proposed algorithm with varying the number of iterations k and the values of c_0. (a) SNR ≈ 7.1; (b) SNR ≈ 3.6.

Simulation I. Besides the number of iterations, there are two additional parameters, the re-scaling parameter α_k and the step-size parameter l_k, may play important roles on the learning performance of KReBooT. According to our theoretical assertions, if $\alpha_k = \frac{2}{k+2}$ and $l_k \sim c_0 \log(k+1)$ for some $c_0 > 0$, then KReBooT can attain a fast learning rate. Thus, this simulation mainly focuses on investigating the effect of the constant c_0 on the prediction performance of KReBooT. To this end, we generate $m = 1000$ samples for training and $m' = 2000$ samples for testing, under two noise levels in terms of signal to noise ratio (SNR). That is, we simulated noise $\{\varepsilon_i\}_{i=1}^N$ independently from $N(0, \sigma^2)$ and we vary σ^2 to obtain SNR ≈ 7.1 and SNR ≈ 3.6, respectively. We then consider 50 candidates of c_0 that logarithmical equally spaced in $[0.1, 80]$.

Figure 2 gives the visualization of testing mean-squared errors (MSE) (in the logscale) of KReBooT with $\alpha_k = \frac{2}{k+2}$ via varying the number of iterations k and the value of c_0. For any fixed step of iteration and c_0, the testing MSE is the average result over 20 independent trails. It is easy to observe from this figure that, there exits a number of c_0's in $[0.1, 5]$ for both noise level cases, such that the testing MSE attains a stable value, neglecting the increasing of iterations. This finding means that an appropriate choice of c_0 could avoid over-fitting. Therefore, for simplicity, we fix $c_0 = 0.5$ in the following simulations.

Simulation II. The objective of this simulation is to describe the relation between the prediction accuracy and the size of training samples for the proposed KReBooT. We thus generate $m = 300, 900, 1500, 4500, 12000$ samples, respectively, for training, and $m' = 2000$ samples for testing. Similar to the previous Simulation I, we also consider two noise levels with SNR ≈ 7.1 and SNR ≈ 3.6, respectively.

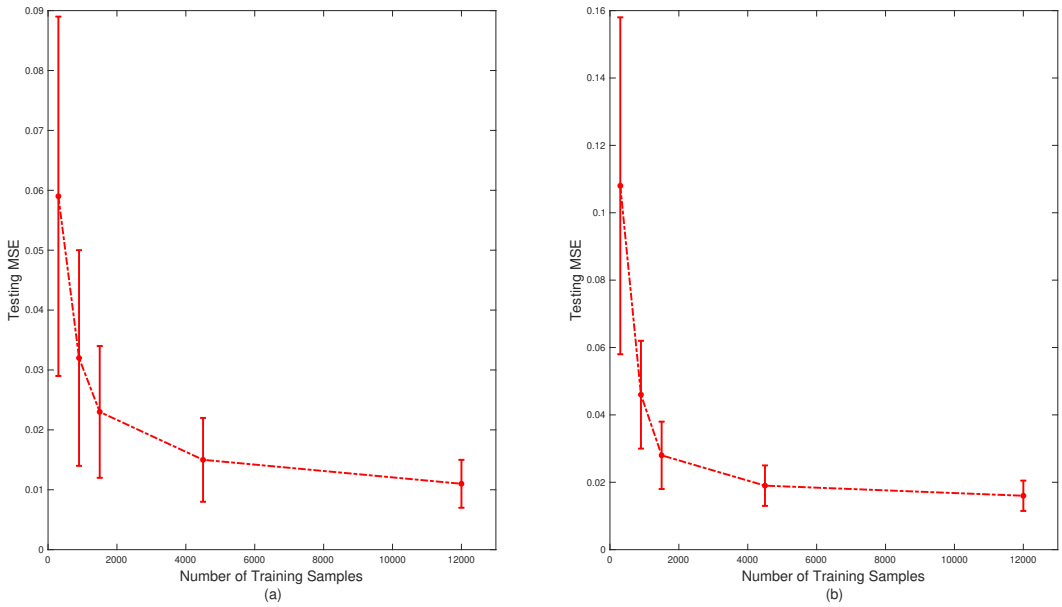

Figure 3: The testing MSE versus the different sizes of training samples. (a) SNR $\approx$ 7.1; (b) SNR $\approx$ 3.6.

Figure 3 depicts the average results over 100 independent trials. It is not hard to observe from this figure that the testing MSE decreases as the number of training samples increases in two noise levels.

Simulation III. In this simulation, we shall compare the prediction performance of the proposed KReBooT with one tree based boosting method, that is, gradient boosted decision trees (GBDT) (Friedman, 2001) (here the number of tree split is 4), and several popular kernel-based methods, including kernel Lasso (Klasso) (Wang et al., 2007), kernel ridge regression with Nyström approximation (KRRN) (Grittens and Mahoney, 2016), and three kernel version of popular boosting algorithms, i.e., ϵ-boosting (Hastie et al., 2007), rescale-boosting (Wang et al., 2019), regularized boosting with truncation (Zhang and Yu, 2005). We refer to these three kernel-based boosting algorithms as ϵ-Kboosting, KRboosting and KRTboosting, respectively. For two different noise levels, we firstly generate $m = 500$, 1000, and 10000 samples to built up the training set, and then generate a validation set of size 1000 for tuning the parameters of different methods, and another 2000 samples to evaluate the performances in terms of MSE. The hyperparameter tuning procedure of KReBooT is the same to the previous simulations, while those of the other methods are the default procedures stated in the aforementioned references.

Table 3 documents the average MSE over 100 independent runs. Numbers in parentheses are the standard errors. It is not hard to see that the performance of the proposed KReBooT is a little bit worse than GBTD, and comparable with Klasso and KRboosting, and clearly better than others. Several important things should be further emphasized. Firstly, though KRboosting illustrates a similar good generalization capability as KReBooT, this algorithm is more likely to overfit, just as the following simulation shown. Secondly, Klasso

Table 3: Prediction Performance of Different Methods under Different Settings

Training Size	Noise Level	Methods						
		KReBooT	GBDT	KRboosting	KRTboosting	ϵ-Kboosting	Klasso	KRRN
$m = 500$	SNR ≈ 7.1	0.059 ± 0.034	0.056 ± 0.031	0.059 ± 0.035	0.073 ± 0.036	0.073 ± 0.035	0.060 ± 0.033	0.092 ± 0.036
	SNR ≈ 3.6	0.088 ± 0.039	0.085 ± 0.035	0.089 ± 0.042	0.105 ± 0.049	0.102 ± 0.048	0.088 ± 0.036	0.127 ± 0.047
$m = 1000$	SNR ≈ 7.1	0.030 ± 0.013	0.026 ± 0.012	0.029 ± 0.011	0.034 ± 0.012	0.032 ± 0.010	0.029 ± 0.009	0.045 ± 0.013
	SNR ≈ 3.6	0.043 ± 0.019	0.039 ± 0.020	0.043 ± 0.019	0.048 ± 0.020	0.047 ± 0.021	0.042 ± 0.019	0.071 ± 0.023
$m = 10000$	SNR ≈ 7.1	0.014 ± 0.005	0.012 ± 0.005	0.015 ± 0.006	0.019 ± 0.008	0.018 ± 0.006	0.013 ± 0.007	0.023 ± 0.009
	SNR ≈ 3.6	0.021 ± 0.010	0.018 ± 0.010	0.021 ± 0.011	0.027 ± 0.013	0.027 ± 0.011	0.020 ± 0.009	0.033 ± 0.012

requires much more time in the training process than the proposed KReBooT. Thirdly, GBDT lacks similar good theoretical guarantees as KReBooT, although its performance is mostly slight better than KReBooT.

Simulation IV. In this simulation, we shall show the overfitting resistence of KReBooT, as compared with its two cousins, i.e., KRboosting and KRTboosting. Here we generate 1000 samples for training, and another 2000 samples for testing, under the noise level of SNR ≈ 7.1. Figure 4(a) clearly demonstrates the merits of our proposed KReBooT, that is, the obtained testing MSE doesn't increase as the number of iterations increases. In addition, different from KRboosting and KRTboosting, the ℓ_1 norm of the coefficients obtained by KReBooT could converge to a fixed value as shown in Figure 4(b), which conforms to the assertion of structure constraints of KReBoot, just as Lemma 1 purports to show.

Simulation V. In this simulation, we mainly show the coefficients estimation behavior of KReBooT. Similar to the above simulation, KRboosting and KRTboosting are considered for comparison. The aim now is to show the reason why KReBooT is overfitting-resistant. Thus, we generate 1000 samples for training, and another 2000 samples for testing, under two different noise levels, that is, SNR $\approx$ 7.1 and SNR $\approx$ 3.6. Figures 5 exhibits the numerical results. It can be found in both cases that the ℓ_1 norm of the KReBooT estimator increases much slower than that of other algorithms, which implies that the variance of KReBooT keeps almost the same and is much smaller than other algorithms when the iterations increases. Thus, the generalization error does not increase very much as the iteration happens.

5.2 Real data experiments

In this subsection, we shall compare the performance of KReBooT over several other alternatives on three commonly-used real data sets. The first one is the Diabetes data set (Efron et al., 2004), which contains 442 diabetes patients that are measured on ten independent variables, that is, age, sex, body mass index etc., and one response variable, that is, a measure of disease progression. The second one is the Concrete Compressive Strength (CCS) data set (Ye, 1998), which contains 1030 instances including eight quantitative independent variables, that is, age, ingredients etc., and one dependent variable, that is, quantitative concrete compressive strength. The third one is the Abalone data set, which comes from an original study in (Nash et al., 1994) for predicting the age of abalone from physical measurements. The data set contains 4177 instances which were measured on eight independent variables, i.e., length, sex, height etc., and one response variable, i.e., the number of rings.

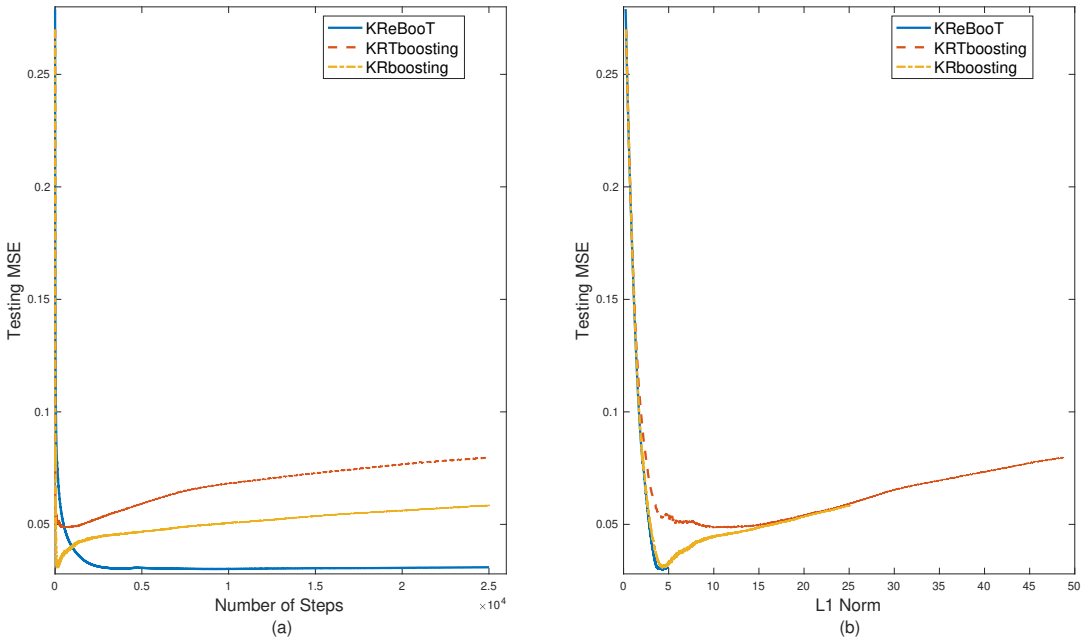

Figure 4: The prediction results obtained by three different kernel-based boosting algorithms. (a) The testing MSE versus the number of steps; (b) The testing MSE versus the ℓ_1 norm.

We use the standard Gaussian kernel for all the kernel based methods, and CART with 4 splits as week learners for GBDT. For each data set, we randomly (according to the uniform distribution) select 50% data for training, 25% data to build the validation set for tuning the parameters and the remainder 25% data as the test set for evaluating the performances of different algorithms. Table 4 documents the performance of all the compared methods. Here we use the average RMSE (root mean squared error) over 100 independent runs as the performance measure. Numbers in parentheses are the standard errors. The hyperparameter tuning strategies of all boosting-type algorithms are the same as those in the previous simulations. It can be easily observed that the performance of KReBooT is a little bit worse than GBDT, and comparable with KRboosting and Klasso, and clearly better than other three methods. These observations coincide with the previous simulations and therefore, experimentally show the merits of the proposed KReBooT.

Table 4: Prediction Performance (numbers in parentheses are the standard errors) of Different Methods on Three Real Data Sets

Data Sets	Methods						
	KReBooT	GBDT	KRboosting	KRTboosting	ϵ-Kboosting	Klasso	KRRN
Diabetes	56.66 ± 3.76	56.32 ± 3.62	56.79 ± 4.54	57.84 ± 3.64	57.95 ± 3.45	56.2 ± 3.86	58.99 ± 4.61
CCC	5.27 ± 0.178	5.23 ± 0.167	5.33 ± 0.185	5.56 ± 0.191	5.86 ± 0.195	5.30 ± 0.169	5.84 ± 0.173
Abalone	2.21 ± 0.071	2.19 ± 0.051	2.18 ± 0.054	2.35 ± 0.075	2.26 ± 0.048	2.22 ± 0.049	2.37 ± 0.067

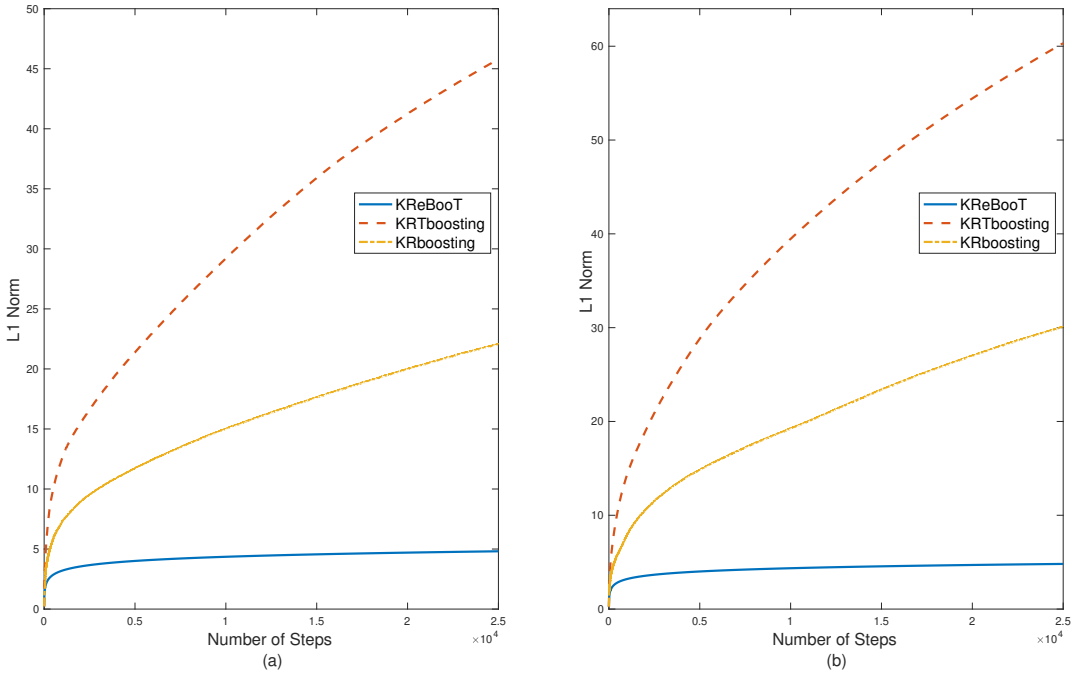

Figure 5: The coefficients estimated by three different kernel-based boosting algorithms versus the number of steps. (a) SNR ≈ 7.1; (b) SNR ≈ 3.6.

6. Proofs

In this section, we present the detailed proofs of our main results. Section 6.1 aims at proving Theorem 2. Section 6.2 devotes to proving Theorem 3. Section 6.3 focus on proving Theorem 8. Section 6.4 refers to proving Proposition 7 and Corollary 9. In Section 6.5, we provide a proof for the concentration inequality (Theorem 14 below) that is used in Section 6.3.

6.1 Proof of Theorem 2

Before presenting the proof of Theorem 2, we at first prove the following lemma to show the role of iterations in KReBooT.

Lemma 11 *Let* $\Lambda_k := [-\alpha_k l_k, \alpha_k l_k]$. *For an arbitrary* $h \in \mathcal{H}_{K,D}$ *with* $\|h\|_{\ell_1} < \infty$, *if* $|y_i| \le M$, $\kappa \le 1$ *and* $\{l_k\}$ *is non-decreasing, then*

$$\|y - f_{D,k}\|_m^2 - \|y - h\|_m^2 \le (1 - \alpha_k)(\|y - f_{D,k-1}\|_m^2 - \|y - h\|_m^2) + \alpha_k^2 (M + \|h\|_{\ell_1})^2 \quad (20)$$

holds for all $k \ge k_h^* := \arg\min_k \{l_k \ge \|h\|_{\ell_1}\}$.

Proof For $k \ge k_h^*$, denote $\beta = \alpha_k \|h\|_{\ell_1} \mathrm{sign}\langle y - f_{D,k-1}, g_k^* \rangle_m$. By the definition of g_k^*, for any $u \in \pm S$ (that is, $u \in S$ or $-u \in S$),

$$\langle y - f_{D,k-1}, \beta g_k^* \rangle_m \ge \alpha_k \|h\|_{\ell_1} \langle y - f_{D,k-1}, u \rangle_m. \quad (21)$$

It is easy to see that (21) also holds for all u on the convex hull B_1 of $\pm S$. So we let $u = h/\|h\|_{\ell_1}$ to obtain

$$\langle y - f_{D,k-1}, \beta g_k^* \rangle_m \geq \alpha_k \langle y - f_{D,k-1}, h \rangle_m.$$

We have

$$\begin{aligned} & \|y - f_{D,k}\|_m^2 \leq \|(1-\alpha_k)y + \alpha_k y - (1-\alpha_k) f_{D,k-1} - \beta g_k^*\|_m^2 \\ \leq \; & (1-\alpha_k)^2 \|y - f_{D,k-1}\|_m^2 + \|\alpha_k y - \beta g_k^*\|_m^2 + 2(1-\alpha_k)\langle y - f_{D,k-1}, \alpha_k y - \alpha_k h \rangle_m. \end{aligned}$$

Now $|y_i| \leq M$ and $h \in B_{l_k}$ yields

$$\|\alpha_k y - \beta g_k^*\|_m^2 \leq \alpha_k^2 (M + \|h\|_{\ell_1})^2.$$

Therefore,

$$\begin{aligned} & \|y - f_{D,k}\|_m^2 - \|y - h\|_m^2 - (1-\alpha_k)\|y - f_{D,k-1}\|_m^2 + (1-\alpha_k)\|y - h\|_m^2 \\ \leq \; & [(1-\alpha_k)^2 - (1-\alpha_k)]\|y - f_{D,k-1}\|_m^2 + \|\alpha_k y - \beta g_k^*\|_m^2 \\ + \; & 2(1-\alpha_k)\alpha_k \langle y - f_{D,k-1}, y - h \rangle_m - \alpha_k \|y - h\|_m^2 \\ \leq \; & [-\alpha_k(1-\alpha_k) + \alpha_k(1-\alpha_k)]\|y - f_{D,k-1}\|_m^2 + \alpha_k^2 (M + \|h\|_{\ell_1})^2 - \alpha_k^2 \|y - h\|_m^2 \\ \leq \; & \alpha_k^2 (M + \|h\|_{\ell_1})^2. \end{aligned}$$

This completes the proof of Lemma 11. ■

Lemma 12 *Let $\{\eta_k\}_{k=0}^{\infty}$ be a sequence of nonnegative numbers, $C \geq 0$ be a constant, and $\{\alpha_k\}_{k=1}^{\infty} \subset (0,1)$ be a nonincreasing sequence of numbers with*

$$\lim_{k\to\infty} \alpha_k = 0, \quad \textit{and} \quad \sum_{k=1}^{\infty} \alpha_k = \infty. \tag{22}$$

If

$$\eta_k \leq (1-\alpha_k)\eta_{k-1} + \alpha_k^2 C, \quad \textit{for all } k \geq 1, \tag{23}$$

then $\lim_{k\to\infty} \eta_k = 0$.

Proof

For any $k \geq 1$, if $\eta_k > \eta_{k-1}$, then (23) implies that

$$\eta_{k-1} \leq (1-\alpha_k)\eta_{k-1} + \alpha_k^2 C,$$

so $\eta_{k-1} \leq \alpha_k C$. We substitute this back to (23) to obtain $\eta_k \leq (1-\alpha_k)\alpha_k C + \alpha_k^2 C = \alpha_k C$. Therefore, $\eta_k \leq \max\{\eta_{k-1}, \alpha_k C\}$. Since $\{\alpha_k\}$ is nonincreasing,

$$\max\{\eta_k, \alpha_{k+1} C\} \leq \max\{\eta_{k-1}, \alpha_k C\}, \quad \text{for all } k \geq 1.$$

Denote $\tilde{\eta}_k = \max\{\eta_k, \alpha_{k+1}C\}$. Then $\{\tilde{\eta}_k\}$ is a nonincreasing sequence of nonnegative numbers, so the limit $\eta^* := \lim_{k\to\infty} \tilde{\eta}_k \geq 0$ exists and is finite.

If $\eta^* > 0$, since $\alpha_k \downarrow 0$, there exists some k^* such that $\eta_k = \tilde{\eta}_k \downarrow \eta^* > 0$ and $\alpha_k C \leq \frac{1}{2}\eta^*$, for all $k \geq k^*$. Recall that $1 - t \leq e^{-t}$ for all $t \in \mathbb{R}$. Let $n > k \geq k^*$ to have

$$\begin{aligned} 0 \leq \eta_n &\leq (1-\alpha_n)\eta_{n-1} + \alpha_n^2 C \leq (1-\alpha_n)\eta_{n-1} + \frac{1}{2}\alpha_n\eta_{n-1} \\ &\leq \cdots \leq \eta_k \prod_{j=k+1}^{n} \left(1 - \frac{\alpha_j}{2}\right) \leq \eta_k \exp\left\{-\frac{1}{2}\sum_{j=k+1}^{n}\alpha_j\right\} \to 0, \quad \text{as } n \to \infty, \end{aligned}$$

where the last limit follows (22). This contradicts $\eta^* > 0$. Therefore $\eta^* = 0$, and the proof is complete. ■

With the help of the above lemmas, we are in a position to prove Theorem 2.

Proof [Proof of Theorem 2] Write

$$h_\infty = \begin{cases} \arg\min\limits_{f\in\mathcal{H}_{K,D}} \|f - y\|_m^2, & \text{when } l_k \uparrow \infty, \\ \arg\min\limits_{f\in B_L} \|f - y\|_m^2, & \text{when } l_k \equiv L. \end{cases}$$

Then there is some integer k^* such that $l_k \geq \|h_\infty\|_{\ell_1}$ whenever $k \geq k^*$. In particular, $k^* = \arg\min_k\{l_k \geq \|h_\infty\|_{\ell_1}\}$ when $l_k \uparrow \infty$, and $k^* = 1$ when $l_k \equiv L$. When $k \geq k^*$, Lemma 11 implies that

$$\|y - f_{D,k}\|_m^2 - \|y - h_\infty\|_m^2 \leq (1-\alpha_k)\left(\|y - f_{D,k-1}\|_m^2 - \|y - h_\infty\|_m^2\right) + \alpha_k^2(M + \|h_\infty\|_{\ell_1})^2. \tag{24}$$

In particular, in the setting $l_K = L$, Lemma 1 guarantees $f_{D,k} \in B_L$ for all k. Therefore the left-hand side of (24) is nonnegative for both the case $l_k \uparrow \infty$ and the case $l_k \equiv L$. We apply Lemma 12 with $\eta_k = \|y - f_{D,k}\|_m^2 - \|y - h_\infty\|_m^2$ and $C = (M + \|h_\infty\|_{\ell_1})^2$ to obtain

$$\lim_{k\to\infty} \|y - f_{D,k}\|_m^2 = \|y - h_\infty\|_m^2. \tag{25}$$

The argument to prove $f_{D,k} \to h_\infty$ is standard. In fact, since h_∞ is the minimizer of $\|f - y\|_m^2$ on $\mathcal{H}_{K,D}$ (rsp., B_L),

$$\langle y - h_\infty, f - h_\infty\rangle_m \leq 0, \tag{26}$$

for any $f \in \mathcal{H}_{K,D}$ (rsp., $f \in B_L$), otherwise $\|h_\infty + t(f - h_\infty) - y\|_m^2 = \|h_\infty - y\|_m^2 + 2t\langle h_\infty - y, f - h_\infty\rangle_m + t^2\|f - h_\infty\|_m^2 < \|h_\infty - y\|_m^2$ for sufficiently small $t > 0$, contradicting the assumption that h_∞ is the minimizer. Now we apply the inequality (26) to obtain

$$\begin{aligned} &\|y - f_{D,k}\|_m^2 - \|y - h_\infty\|_m^2 \\ =&\|f_{D,k}\|_m^2 - 2\langle y - h_\infty, f_{D,k} - h_\infty\rangle_m - 2\langle h_\infty, f_{D,k} - h_\infty\rangle_m - \|h_\infty\|_m^2 \\ \geq&\|f_{D,k} - h_\infty\|_m^2 \to 0, \quad \text{as } k \to \infty. \end{aligned}$$

The proof is complete. ■

6.2 Proof of Theorem 3

Proof [Proof of Theorem 3] From the definition (3) of iteration, we have for all $k \geq 1$ that

$$\|y - f_{D,k}\|_m \leq \|(1-\alpha_k) f_{D,k-1} - y\|_m \leq \alpha_k \|y\|_m + (1-\alpha_k) \|y - f_{D,k-1}\|_m . \tag{27}$$

From the assumption $|y_i| \leq M$ we have $\|y\|_m \leq M$. Starting from $f_{D,0} = 0$, we use (27) iteratively to have

$$\|y - f_{D,k}\|_m \leq M, \quad \text{for all } k \geq 1. \tag{28}$$

Denote

$$A_k = \|y - f_{D,k}\|_m^2 - \|y - h\|_m^2 , \quad k \geq 1.$$

From the setting $\alpha_k = \frac{2}{k+2}$ and Lemma 11, we have

$$A_k \leq \left(1 - \frac{2}{k+2}\right) A_{k-1} + \frac{4(M + \|h\|_{\ell_1})^2}{(k+2)^2}, \quad \text{for all } k \geq k_h^*. \tag{29}$$

In particular, (28) implies $A_k \leq M^2$ for all k. We use (29) iteratively to obtain that for $k \geq k_h^*$,

$$\begin{aligned}
A_k &\leq \frac{k}{k+2}\left(\frac{k-1}{k+1} A_{k-2} + \frac{4(M + \|h\|_{\ell_1})^2}{(k+1)^2}\right) + \frac{4(M + \|h\|_{\ell_1})^2}{(k+2)^2} \\
&\leq \cdots \leq A_{k_h^*} \prod_{i=k_h^*+1}^{k} \frac{i}{i+2} + \sum_{i=k_h^*+1}^{k} \frac{4(M + \|h\|_{\ell_1})^2}{(i+2)^2} \prod_{j=i+1}^{k} \frac{j}{j+2} \\
&= \frac{(k_h^*+1)(k_h^*+2)}{(k+2)(k+1)} A_{k_h^*} + \sum_{i=k_h^*+1}^{k} \frac{4(M + \|h\|_{\ell_1})^2}{(i+2)^2} \frac{(i+2)(i+1)}{(k+2)(k+1)} \\
&\leq \frac{(k_h^*+1)(k_h^*+2) M^2}{(k+1)(k+2)} + \frac{4(M + \|h\|_{\ell_1})^2 (k - k_h^*)}{(k+1)(k+2)},
\end{aligned}$$

therefore,

$$A_k \leq \left[M^2 (k_h^*+2)^2 + 4(M + \|h\|_{\ell_1})^2\right] / (k+2). \tag{30}$$

Meanwhile, when $k < k_h^*$, from $A_k \leq M^2$ one immediately concludes (30). This completes the proof. ■

6.3 Proof of Theorem 8

We adopt a standard error decomposition technique in learning theory (Shi et al., 2011) to Theorem 8 and get the following proposition directly.

Proposition 13 *Let $\delta \in (0,1)$. If Assumption 2 holds, then with confidence $1 - \delta/2$, there holds*

$$\begin{aligned}
\mathcal{E}(f_{D,k}) - \mathcal{E}(f_\rho) &\leq c_1(\delta/2) m^{-2r} + 32 \max\left\{16 k_{\delta/2}^* \left[M^2 (k_{\delta/2}^* + 4) + 8 l_{k_{\delta/2}^*}^2\right], 15\right\} k^{-1} \\
&+ \mathcal{S}_1(D,k) + \mathcal{S}_2(D),
\end{aligned}$$

where

$$\mathcal{S}_1(D,k) := [\mathcal{E}(f_{D,k}) - \mathcal{E}(f_\rho)] - [\mathcal{E}_D(f_{D,k}) - \mathcal{E}_D(f_\rho)], \tag{31}$$

$$\mathcal{S}_2(D) := [\mathcal{E}_D(h_{D,\rho}) - \mathcal{E}_D(f_\rho)] - [\mathcal{E}(h_{D,\rho}) - \mathcal{E}(f_\rho)] \tag{32}$$

and $\mathcal{E}_D(f) := \frac{1}{m}(f(x_i) - y_i)^2$.

Proof Denoting $\mathcal{A}(D) := \mathcal{E}(h_{D,\rho}) - \mathcal{E}(f_\rho)$ and $\mathcal{H}(D,k) := \mathcal{E}_D(f_{D,k}) - \mathcal{E}_D(h_{D,\rho})$, we have

$$\mathcal{E}(f_{D,k}) - \mathcal{E}(f_\rho) = \mathcal{S}_1(D,k) + \mathcal{S}_2(D) + \mathcal{H}(D,k) + \mathcal{A}(D).$$

Setting $h_D = h_{D,\rho}$, we then get from Assumption 2 with $\tau = \delta/2$ and (9) that with confidence $1 - \delta/2$, there holds

$$\mathcal{A}(D) \le c_1(\delta/2)m^{-2r}.$$

Furthermore, it follows from Theorem 3 and Assumption 2 that

$$\mathcal{H}(D,k) \le 32 \max\left\{16k^*_{\delta/2}\left[M^2(k^*_{\delta/2}+4) + 8l^2_{k^*_{\delta/2}}\right], 15\right\} k^{-1}.$$

Combining the above three estimates, we complete the proof of Proposition 13. ■

Based on Proposition 13, the key step-stone to provide a tight generalization error of $f_{D,k}$ is to bound the sample error $\mathcal{S}_1(D,k) + \mathcal{S}_2(D)$. Recalling (31) and (32), both $h_{D,\rho}$ and $f_{D,k}$ depend on the sampling, making the classical concentration inequality approaches in (Zhang and Yu, 2005; Shi et al., 2011; Shi, 2013; Wang et al., 2019) infeasible since all these approaches require $h_{D,\rho}$ to be independent of D. Under this circumstance, a novel concentration inequality is needed. To this end, we derive the following concentration inequality, which is a modified version of (Steinwart and Christmann, 2008, Theorem 7.20). We present its proof in Section 6.5.

Theorem 14 *Let $0 < \delta < 1$ and $R > 0$. If $|y_i| \le M$, $\kappa \le 1$ and Assumption 1 holds with some $c > 0$ and $0 < s < 1$, then with confidence $1 - \delta$, there holds*

$$|\{\mathcal{E}(f) - \mathcal{E}(f_\rho)\} - \{\mathcal{E}_D(f) - \mathcal{E}_D(f_\rho)\}| \le \frac{1}{2}(\mathcal{E}(f) - \mathcal{E}(f_\rho)) + \frac{32(3M+R)^2 \log\frac{1}{\delta}}{3m}$$
$$+ \quad \bar{C}(3M+R)^2 \max\left\{(\mathcal{E}(f) - \mathcal{E}(f_\rho))^{\frac{1-s}{2}} m^{-\frac{1}{2}}, m^{-\frac{1}{1+s}}\right\}, \qquad \forall f \in B_{K,R}, \tag{33}$$

where $B_{K,R} := \{f \in \mathcal{H}_K : \|f\|_K \le R\}$ and $\bar{C}$ is a constant depending only on s and c.

We then use the concentration inequality established in Theorem 14 to derive the upper bound of $\mathcal{S}_2(D,\lambda)$.

Proposition 15 *Let $0 < \delta < 1$, $|y_i| \le M$ and $\kappa \le 1$. If Assumptions 2 and 1 hold with some $c > 0$ and $0 < s < 1$, then with confidence at least $1 - \delta$, there holds*

$$\mathcal{S}_2(D) \le C_1(c_1(\delta/2) + (c_2(\delta/2))^2(c_1(\delta/2))^{(1-s)/2}) \max\left\{m^{-\frac{1}{1+s}}, m^{-(1-s)r-1/2}\right\} \log\frac{2}{\delta}, \tag{34}$$

where C_1 is a constant depending only on M, s, c.

Proof Due to (9) and Assumption 2 with $\tau = \delta/2$, with confidence $1 - \delta/2$, there holds

$$\mathcal{E}(h_{D,\rho}) - \mathcal{E}(f_\rho) \leq c_1(\delta/2)m^{-2r}, \qquad \text{and} \|h_{D,\rho}\|_{\ell_1} \leq c_2(\delta/2).$$

Then, it follows from Theorem 14 with $R = c_2(\delta/2)$ that with confidence $1 - \delta$, there holds

$$\begin{aligned}\mathcal{S}_2(D) &\leq \frac{c_1(\delta/2)}{2}m^{-2r} + \frac{32(3M + c_2(\delta/2))^2 \log\frac{2}{\delta}}{3m} \\ &+ \bar{C}(3M + c_2(\delta/2))^2 \max\left\{(c_1(\delta/2)m^{-2r})^{\frac{1-s}{2}}m^{-\frac{1}{2}}, m^{-\frac{1}{1+s}}\right\} \\ &\leq C_1(c_1(\delta/2) + (c_2(\delta/2))^2(c_1(\delta/2))^{(1-s)/2}) \max\left\{m^{-\frac{1}{1+s}}, m^{-(1-s)r-1/2}\right\}\log\frac{2}{\delta},\end{aligned}$$

where C_1 is a constant depending only on $M, \bar{C}$. This completes the proof of Proposition 15. ■

In the following, we aim to derive the estimate for $\mathcal{S}_1(D, k)$.

Proposition 16 *Let $0 < \delta < 1$, $|y_i| \leq M$, $\kappa \leq 1$ and $f_{D,k}$ be defined by (2) with $\alpha_k = \frac{2}{k+2}$ and $l_k = c_0 \log(k+1)$. If Assumptions 1 holds with some $c > 0$ and $0 < s < 1$ and*

$$\mathcal{E}(f_{D,k}) - \mathcal{E}(f_\rho) \geq m^{-\frac{1}{1+s}}, \tag{35}$$

then

$$\begin{aligned}\mathcal{S}_1(D, k) &\leq \frac{1}{2}(\mathcal{E}(f_{D,k}) - \mathcal{E}(f_\rho)) + \frac{32(3M + c_0\log(k+1))^2 \log\frac{2}{\delta}}{3m} \\ &+ \bar{C}(3M + c_0\log(k+1))^2(\mathcal{E}(f_{D,k}) - \mathcal{E}(f_\rho))^{\frac{1-s}{2}}m^{-\frac{1}{2}}\end{aligned}$$

holds with confidence $1 - \delta$.

Proof Due to (35), there holds

$$\max\left\{(\mathcal{E}(f_{D,k}) - \mathcal{E}(f_\rho))^{\frac{1-s}{2}}m^{-\frac{1}{2}}, m^{-\frac{1}{1+s}}\right\} = (\mathcal{E}(f_{D,k}) - \mathcal{E}(f_\rho))^{\frac{1-s}{2}}m^{-\frac{1}{2}}. \tag{36}$$

But Theorem 3 together with $\kappa \leq 1$ implies $f_{D,k} \in B_R \subset B_{K,R}$ with $R = c_0 \log(k+1)$. Then it follows from Theorem 14 that with confidence $1 - \delta$ there holds

$$\begin{aligned}\mathcal{S}_1(D, k) &\leq \frac{1}{2}(\mathcal{E}(f_{D,k}) - \mathcal{E}(f_\rho)) + \frac{32(3M + c_0\log(k+1))^2 \log\frac{1}{\delta}}{3m} \\ &+ \bar{C}(3M + c_0\log(k+1))^2(\mathcal{E}(f_{D,k}) - \mathcal{E}(f_\rho))^{\frac{1-s}{2}}m^{-\frac{1}{2}}.\end{aligned}$$

This completes the proof of Proposition 16. ■

With the help of the above three propositions, we are in a position to prove Theorem 8 as follows.

Proof of Theorem 8. If (35) does not hold, then we obtain (15) directly. If (35) holds, it follows from Propositions 13, 15 and 16 by scaling δ to $\delta/3$ that with confidence $1-\delta$, there holds

$$\begin{aligned}
&\frac{1}{2}(\mathcal{E}(f_{D,k})-\mathcal{E}(f_\rho)) \le c_1(\delta/6)m^{-2r} + 32\max\left\{16k^*_{\delta/6}\left[M^2(k^*_{\delta/6}+4)+8l^2_{k^*_{\delta/6}}\right], 15\right\}k^{-1}\\
+\quad & C_1(c_1(\delta/6)+(c_2(\delta/6))^2(c_1(\delta/6))^{(1-s)/2})\max\left\{m^{-\frac{1}{1+s}}, m^{-(1-s)r-1/2}\right\}\log\frac{6}{\delta}\\
+\quad & \frac{32(3M+c_0\log(k+1))^2\log\frac{6}{\delta}}{3m} + \bar{C}(3M+c_0\log(k+1))^2(\mathcal{E}(f_{D,k})-\mathcal{E}(f_\rho))^{\frac{1-s}{2}}m^{-\frac{1}{2}}.
\end{aligned}$$

Since $m^{-\frac{1}{1+s}} = \left(m^{-\frac{1}{1+s}}\right)^{\frac{1-s}{2}} m^{-\frac{1}{2}}$, $k \ge \min\left\{m^{\frac{1}{1+s}}, m^{(1-s)r+1/2}\right\}$ and (35) holds, we have

$$\mathcal{E}(f_{D,k})-\mathcal{E}(f_\rho) \le C_1(\delta)\log^2(k+1)\max\left\{(\mathcal{E}(f_{D,k})-\mathcal{E}(f_\rho))^{\frac{1-s}{2}}m^{-1/2}, m^{-(1-s)r-1/2}\right\}\log\frac{6}{\delta},$$

where

$$C_1(\delta) := C_1'(c_1(\delta/6)+k^*_{\delta/6}(k^*_{\delta/6}+l^2_{k^*_{\delta/6}})+(c_2(\delta/6))^2(c_1(\delta/6))^{(1-s)/2}),$$

where C_1' is a constant depending only on M, s, c. Hence, with confidence $1-\delta$, there holds

$$\mathcal{E}(f_{D,k})-\mathcal{E}(f_\rho) \le (C_1(\delta))^2(\log(k+1))^4\max\left\{m^{-\frac{1}{1+s}}, m^{-(1-s)r-1/2}\right\}\log^2\frac{6}{\delta},$$

This proves Theorem 8 with $C(\delta) := (C_1(\delta))^2$. ■

6.4 Proofs of Proposition 7 and Corollary 9

For an arbitrary $\lambda > 0$ define

$$f^0_\lambda = (L_K+\lambda I)^{-1}f_\rho,\ f_\lambda = (L_K+\lambda I)^{-1}L_K f_\rho,\ f^0_{D,\lambda} = L_{K,D}(L_K+\lambda I)^{-1}f_\rho, \tag{37}$$

where $L_{K,D}: \mathcal{H}_K \to \mathcal{H}_K$ is the empirical operator defined by

$$L_{K,D}f := \frac{1}{m}\sum_{i=1}^m f(x_i)K_{x_i}.$$

We will show that $f^0_{D,\lambda}$ is the desired $h_{D,\rho}$ in Assumption 2. For this purpose, we need two important lemmas. The first one is the well known Bernstein inequality (Shi et al., 2011).

Lemma 17 *Let ξ be a random variable on $\mathcal{Z}$ with variance γ^2 satisfying $|\xi - E[\xi]| \le M_\xi$ for some constant M_ξ. Then for any $0<\delta<1$, with confidence $1-\delta$, we have*

$$\frac{1}{m}\sum_{i=1}^m \xi(z_i) - E[\xi] \le \frac{2M_\xi\log\frac{1}{\delta}}{3m} + \sqrt{\frac{2\gamma^2\log\frac{1}{\delta}}{m}}.$$

The second one focuses on the operator difference, which was derived in (Blanchard and Krämer, 2016; Lin et al., 2017).

Lemma 18 *Let $0 < \delta < 1$. If $\kappa \leq 1$, then with confidence at least $1 - \delta$, there holds*

$$\left\| (L_K + \lambda I)^{-1/2} (L_K - L_{K,D}) \right\| \leq \frac{2}{\sqrt{m}} \left\{ \frac{1}{\sqrt{m\lambda}} + \sqrt{\mathcal{N}(\lambda)} \right\} \log \frac{2}{\delta}, \tag{38}$$

where $\mathcal{N}(\lambda) = Tr\left((L_K + \lambda I)^{-1} L_K\right)$ is the trace of the operator $(L_K + \lambda I)^{-1} L_K$.

With the help of Lemma 17, we derive the following ℓ_1 norm estimate for $f^0_{D,\lambda}$.

Lemma 19 *Let $0 < \delta < 1$ and $\kappa \leq 1$, if there exists an $h_\rho \in L^2_{\rho_X}$ such that $f_\rho = L_K h_\rho$, then with confidence $1 - \delta/2$, there holds*

$$\|f^0_{D,\lambda}\|_{\ell_1} \leq \frac{1}{m} \sum_{i=1}^m |f^0_\lambda(x_i)| \leq \left(1 + \frac{4}{3m\sqrt{\lambda}} \log \frac{2}{\delta} + \sqrt{\frac{2}{m} \log \frac{2}{\delta}} \right) \|h_\rho\|_\rho. \tag{39}$$

Proof We at first bound $\|f^0_{D,\lambda}\|_{\ell_1} - \|f^0_\lambda\|_{L^1_{\rho_X}}$ by using Lemma 17. Let $\xi_1 = |f^0_\lambda(x)|$. We then have

$$E(\xi_1) = \|f^0_\lambda\|_{L^1_{\rho_X}}, \qquad \|f^0_{D,\lambda}\|_{\ell_1} = \frac{1}{m} \sum_{i=1}^m |f^\lambda_0(x_i)| = \frac{1}{m} \sum_{i=1}^m \xi_1(x_i).$$

Due to $\kappa \leq 1$, we have $\|f\|_\infty \leq \|f\|_K$ for arbitrary $f \in \mathcal{H}_K$. Then it follows from $\|h_\rho\|_\rho = \|L_K^{1/2} h_\rho\|_K$, $f^0_\lambda \in \mathcal{H}_K$, (37) and $f_\rho = L_K h_\rho$ that

$$\|f^0_\lambda\|_\infty \leq \|f^0_\lambda\|_K \leq \frac{1}{\sqrt{\lambda}} \|(L_K + \lambda I)^{-1/2} L_K h_\rho\|_K \leq \frac{1}{\sqrt{\lambda}} \|h_\rho\|_\rho. \tag{40}$$

Furthermore,

$$\|f^0_\lambda\|_\rho = \frac{1}{\lambda} \|\lambda (L_K + \lambda I)^{-1} f_\rho\|_\rho = \frac{1}{\lambda} \|f_\lambda - f_\rho\|_\rho \leq \|h_\rho\|_\rho. \tag{41}$$

Then,

$$E[\xi_1^2] = \|f^0_\lambda\|^2_\rho \leq \|h_\rho\|^2_\rho.$$

Hence, for arbitrary $0 < \delta < 1$, Lemma 17 with $\xi = \xi_1 = |f^\lambda_0(x)|$, $M_\xi = \frac{2}{\sqrt{\lambda}} \|h_\rho\|_\rho$ and $\gamma^2 \leq \|h_\rho\|^2_\rho$ yields that with confidence $1 - \delta/2$, there holds

$$\frac{1}{m} \sum_{i=1}^m |f^\lambda_0(x_i)| - \|f^0_\lambda\|_{L^1_{\rho_X}} \leq \left(\frac{4}{3m\sqrt{\lambda}} \log \frac{2}{\delta} + \sqrt{\frac{2}{m} \log \frac{2}{\delta}} \right) \|h_\rho\|_\rho. \tag{42}$$

But (41) implies

$$\|f^0_\lambda\|_{L^1_{\rho_X}} \leq \|f^0_\lambda\|_\rho \leq \|h_\rho\|_\rho. \tag{43}$$

Plugging (43) into (42), with confidence $1 - \delta/2$, there holds

$$\|f^0_{D,\lambda}\|_{\ell_1} \leq \frac{1}{m} \sum_{i=1}^m |f^0_\lambda(x_i)| \leq \left(1 + \frac{4}{3m\sqrt{\lambda}} \log \frac{2}{\delta} + \sqrt{\frac{2}{m} \log \frac{2}{\delta}} \right) \|h_\rho\|_\rho.$$

This completes the proof of Lemma 19. ■

We then derive $\|f^0_{D,\lambda} - \rho\|_\rho$ by using Lemma 18.

Lemma 20 *Let $0 < \delta < 1$ and $\kappa \leq 1$. If Assumption 1 holds with some $c > 0$ and $0 < s < 1$, and there exists an $h_\rho \in L^2_{\rho_X}$ such that $f_\rho = L_K h_\rho$, then*

$$\|f^0_{D,\lambda} - f_\lambda\|_\rho \leq \lambda \|h_\rho\|_\rho + \frac{2}{\sqrt{m}} \left\{ \frac{1}{\sqrt{m\lambda}} + \sqrt{\tilde{c}\lambda^{-s}} \right\} \|h_\rho\|_\rho \log \frac{4}{\delta} \tag{44}$$

holds with confidence $1 - \delta/2$, where $\tilde{c}$ is a constant depending only on c and s.

Proof The triangle inequality yields

$$\|f^0_{D,\lambda} - f_\rho\|_\rho \leq \|f_\lambda - f_\rho\|_\rho + \|f_\lambda - f^0_{D,\lambda}\|_\rho. \tag{45}$$

But $f_\rho = L_K h_\rho$ implies (Caponnetto and De Vito, 2007; Lin et al., 2017)

$$\|f_\lambda - f_\rho\|_\rho \leq \lambda \|h_\rho\|_\rho. \tag{46}$$

Thus, it suffices to bound $\|f_\lambda - f^0_{D,\lambda}\|_\rho$. Due to (37), we have

$$f^0_{D,\lambda} - f_\lambda = (L_{K,D} - L_K)(L_K + \lambda I)^{-1} f_\rho$$

Thus, $\kappa \leq 1$ implies

$$\begin{aligned} & \|f^0_{D,\lambda} - f_\lambda\|_\rho = \|L_K^{1/2}(L_{K,D} - L_K)(L_K + \lambda I)^{-1} L_K h_\rho\|_K \\ \leq \; & \|(L_{K,D} - L_K)(L_K + \lambda I)^{-1/2}\| \|L_K^{1/2} h_\rho\|_K = \|(L_K + \lambda I)^{-1/2}(L_{K,D} - L_K)\| \|h_\rho\|_\rho. \end{aligned}$$

But (11) and the definition of $\mathcal{N}(\lambda)$ yield

$$\begin{aligned} \mathcal{N}(\lambda) &= \sum_{\ell=1}^\infty \frac{\mu_\ell}{\lambda + \mu_\ell} \leq \sum_{\ell=1}^\infty \frac{c\ell^{-1/s}}{\lambda + c\ell^{-1/s}} = \sum_{\ell=1}^\infty \frac{c}{c + \lambda \ell^{1/s}} \\ &\leq \int_0^\infty \frac{c}{c + \lambda t^{1/s}} dt \leq c\lambda^{-s} \left(\int_0^c \frac{1}{c} dt + \int_c^\infty t^{-1/s} dt \right) = \tilde{c}\lambda^{-s}, \end{aligned}$$

where $\tilde{c} := c + \frac{c}{1-s} c^{1-1/s}$. It then follows from Lemma 18 that with confidence at least $1 - \delta/2$, there holds

$$\|f^0_{D,\lambda} - f_\lambda\|_\rho \leq \frac{2}{\sqrt{m}} \left\{ \frac{1}{\sqrt{m\lambda}} + \sqrt{\tilde{c}\lambda^{-s}} \right\} \|h_\rho\|_\rho \log \frac{4}{\delta}.$$

This completes the proof of Lemma 20. ■

Based on the above two lemmas, we can prove Proposition 7 directly.

Proof of Proposition 7. Setting $\lambda = m^{-\frac{1}{s+1}}$, we have from Lemma 19 that with confidence $1 - \delta/2$, there holds

$$\|f^0_{D,\lambda}\|_{\ell_1} \leq 4 \|h_\rho\|_\rho \log \frac{2}{\delta}.$$

Furthermore, it follows from Lemma 20 with $\lambda = m^{-\frac{1}{s+1}}$ that

$$\|f^0_{D,\lambda} - f_\rho\|_\rho \leq (3 + 2\sqrt{\tilde{c}}) m^{-\frac{1}{2(s+1)}} \|h_\rho\|_\rho \log \frac{4}{\delta}.$$

This completes the proof of Proposition 7. ■

We then use Proposition 7 and Theorem 8 to prove Corollary9.

Proof of Corollary 9. It suffices to bound $C(\delta)$ defined in (16) by inserting $c_1(\tau) = \tilde{c}_1\|h_\rho\|_\rho \log\frac{4}{\tau}$ and $c_2(\tau) = 4\|h_\rho\|_\rho \log\frac{2}{\tau}$. Due to the definition of $l_{k^*_{\delta/6}}$ and $l_k = c_0\log(k+1)$, we have

$$k^*_{\delta/6} \le 1 + 4\|h_\rho\|_\rho \log\frac{12}{\delta} \le \bar{C}_1 \log\frac{12}{\delta},$$

and

$$l_{k^*_{\delta/6}} \le \bar{C}_2 \log^{1/2}\frac{12}{\delta},$$

where $\bar{C}_1$ and $\bar{C}_2$ are positive constants depending only on $\|h\|_\rho$ and c_0. Recalling (16), we have

$$C(\delta) \le \bar{C}_3 \log^3\frac{24}{\delta},$$

where $\bar{C}_1$ is a positive constant depending only on $\|h\|_\rho$ and c_0. This proves Theorem 9 by noting $r = \frac{1}{2(s+1)}$ in Theorem 8. ■

6.5 Proof of Theorem 14

Our oracle inequality is built upon the eigenvalue decay assumption (11). We at first connect it with the well known entropy number defined in Definition 21 below.

Definition 21 *Let E be a Banach space and $A \subset E$ be a bounded subset. Then for $i \ge 1$, the i-th entropy number $e_i(A, E)$ of A is the infimum over all $\varepsilon > 0$ for which there exist $t_1, \ldots, t_{2^{i-1}} \in A$ with $A \subset \bigcup_{j=1}^{2^{i-1}}(t_j + \varepsilon B_E)$, where B_E denotes the closed unit ball of E. Moreover, the i-th entropy number of a bounded linear operator $\mathcal{T} : E \to F$ is $e_i(\mathcal{T}) := e_i(\mathcal{T}B_E, F)$, where $\mathcal{T}B_E := \{\mathcal{T}f : f \in B_E\}$.*

We also need the following two lemmas, which can be found in (Steinwart et al., 2009, Theorem 15) and (Steinwart and Christmann, 2008, Corollary 7.31), respectively.

Lemma 22 *Let $\{\mu_i\}_{i=1}^\infty$ be the set eigenvalues of the operator $L_K : L^2_{\rho_X} \to L^2_{\rho_X}$ arranging in a decreasing order. For arbitrary $0 < p < 1$, there exists a constant c'_p depending only on p such that*

$$\sup_{i\le j} i^{\frac{1}{p}} e_i(id : \mathcal{H}_K \to L^2_{\rho_X}) \le c'_p \sup_{i\le j} i^{\frac{1}{p}} \mu_i^{1/2}, \qquad \forall j \ge 1.$$

Lemma 23 *Assume that there exist constants $0 < p < 1$ and $a \ge 1$ such that*

$$e_i(id : \mathcal{H}_K \to L^2_{\rho_X}) \le a i^{-\frac{1}{2p}}, \qquad i \ge 1.$$

Then there exists a constant $c_p > 0$ depending only on p such that

$$E[e_i(id : \mathcal{H}_K \to \ell^2(D_X))] \le c_p i^{-\frac{1}{2p}},$$

where $\ell^2(D_X)$ denotes the empirical ℓ^2 space with respect to $(x_1, \ldots, x_m)$.

With the help of Lemmas 22 and 23, we derive the following upper bound for the empirical entropy number in expectation.

Lemma 24 *If $|y_i| \le M$, $\kappa \le 1$, and Assumption 1 holds with some $c > 0$ and $0 < s < 1$, then*

$$E[e_i(\mathcal{F}_R, \ell^2(D))] \le c_s c_s' \sqrt{c} R(2M + 2R) i^{-\frac{1}{2s}}, \tag{47}$$

where

$$\mathcal{F}_R := \{\phi_f(x) = (f(x) - y)^2 - (f(x) - f_\rho(x))^2 : f \in B_{K,R}\}, \tag{48}$$

and $\ell^2(D)$ denotes the empirical ℓ^2 space with respect to $(z_1, \ldots, z_m)$.

Proof Due to (11) and Lemma 22 with $p = s$, we have for arbitrary $j \ge 1$,

$$j^{\frac{1}{s}} e_j(id : \mathcal{H}_K \to L^2_{\rho_X}) \le \sup_{i \le j} i^{\frac{1}{s}} e_i(id : \mathcal{H}_K \to L^2_{\rho_X}) \le c_s' \sqrt{c} j^{\frac{1}{2s}},$$

which implies

$$e_i(id : \mathcal{H}_K \to L^2_{\rho_X}) \le c_s' \sqrt{c} i^{-\frac{1}{2s}}, \qquad \forall i = 1, 2, \ldots.$$

This together with Lemma 23 yields

$$E[e_i(id : \mathcal{H}_K \to \ell^2(D_X))] \le c_s c_s' \sqrt{c} i^{-\frac{1}{2s}}, \qquad \forall i = 1, 2, \ldots. \tag{49}$$

For arbitrary $f \in B_{K,R}$, there exists an $f^* \in B_{K,1}$ such that $f = Rf^*$. Let $f_1, \ldots, f_{2^{n-1}}$ be an ε net of $B_{K,1}$. Then there exists an f_{j^*} such that

$$\frac{1}{m} \sum_{i=1}^m (f(x_i) - Rf_{j^*}(x_i))^2 = \frac{1}{m} \sum_{i=1}^m (Rf^*(x_i) - Rf_{j^*}(x_i))^2 \le R^2 \varepsilon^2.$$

Thus, $Rf_1, \ldots, Rf_{2^{n-1}}$ is an $R\varepsilon$ net of $B_{K,R}$. This together with (49) implies

$$E[e_i(B_{K,R}, \ell^2(D_X))] \le R E[e_i(B_{K,1}, \ell^2(D_X))] \le c_s c_s' \sqrt{c} R i^{-\frac{1}{2s}}. \tag{50}$$

For arbitrary $\phi_f \in \mathcal{F}_R$, there exists an $f \in B_{K,R}$ such that $\phi_f(x) = (f(x) - y)^2 - (f_\rho(x) - y)^2$. Then there exists an f_{j^*} with $1 \le j^* \le 2^{n-1}$ such that

$$\begin{aligned} & \frac{1}{m} \sum_{i=1}^m (\phi_f(x_i) - ((f_{j^*}(x_i) - y_i)^2 - (f_\rho(x_i) - y_i)^2))^2 \\ = \; & \frac{1}{m} \sum_{i=1}^m (f(x_i) - f_{j^*}(x_i))^2 (f(x_i) + f_{j^*}(x_i) - 2y_i)^2 \le (2M + 2R)^2 R^2 \varepsilon^2, \end{aligned}$$

where we used $|y_i| \le M$ in above estimates. Hence, it follows from (48) and (50) that

$$E[e_i(\mathcal{F}_R, \ell^2(D))] \le (2M + 2R) E[e_i(B_R, \ell^2(D_X))] \le c_s c_s' \sqrt{c} R(2M + 2R) i^{-\frac{1}{2s}}.$$

This completes the proof of Lemma 24. ■

We then present a close relation between the empirical entropy number and empirical Rademacher average (Steinwart and Christmann, 2008, Definitions 7.8&7.9).

Definition 25 *Let $(\Theta, \mathcal{C}, \nu)$ be a probability space and $\epsilon_i : \Theta \to \{-1, 1\}$, $i = 1, \ldots, m$, be independent random variables with $\nu(\epsilon_i = 1) = \nu(\epsilon_i = -1) = 1/2$ for all $i = 1, \ldots, m$. Then, $\epsilon_1, \ldots, \epsilon_m$ is called a Rademacher sequence with respect to ν. Assume $\mathcal{H} \subset \mathcal{M}(\mathcal{Z})$ be a non-empty set with $\mathcal{M}(\mathcal{Z})$ the set of measurable functions on $\mathcal{Z}$. For $D = (z_1, \ldots, z_m) \in \mathcal{Z}^m$, the m-th empirical Rademacher average of $\mathcal{H}$ is defined by*

$$Rad_D(\mathcal{H}, m) := E\left[\sup_{h\in\mathcal{H}}\left|\frac{1}{m}\sum_{i=1}^{m}\epsilon_i h(z_i)\right|\right].$$

The following lemma which were proved in (Steinwart and Christmann, 2008, Lemma 7.6,Theorem 7.16) show that the upper bound of the empirical entropy number of $\mathcal{F}_R$ implies an upper bound of the empirical Rademacher average.

Lemma 26 *Suppose that there exist constants $B_1 \geq 0$ and $\sigma_1 \geq 0$ such that $\|h\|_\infty \leq B_1$ and $E[h^2] \leq \sigma_1^2$ for all $h \in \mathcal{H}$. Furthermore, assume that for a fixed $m \geq 1$ there exist constants $p \in (0, 1)$ and $a \geq B_1$ such that*

$$E[e_i(\mathcal{F}_R, \ell^2(D))] \leq a i^{-\frac{1}{2p}}, \qquad i \geq 1.$$

Then there exist constants C_p and C'_p depending only on p such that

$$E[Rad_D(\mathcal{F}_R, m)] \leq \max\left\{C_p a^p \sigma_1^{1-p} m^{-\frac{1}{2}}, C'_p a^{\frac{2p}{1+p}} B_1^{\frac{1-p}{1+p}} m^{-\frac{1}{1+p}}\right\}.$$

Furthermore, the following lemma proved in (Steinwart and Christmann, 2008, Lemma 7.6,Proposition 7.10), presents the role of the empirical Rademacher average in empirical process.

Lemma 27 *For arbitrary $m \geq 1$ we have*

$$E\left[\sup_{h\in\mathcal{F}_R}\left|E[h] - \frac{1}{m}\sum_{i=1}^{m}h(z_i)\right|\right] \leq 2E[Rad_D(\mathcal{F}_R, m)].$$

Based on the above two lemmas, we can derive the following bound, which plays an important role in our analysis.

Lemma 28 *If $\kappa \leq 1$, $|y_i| \leq M$ and Assumption 1 holds with some $c > 0$ and $0 < s < 1$, then there exists a constant $\widetilde{C}$ depending only on s and c such that*

$$E\left[\sup_{\phi_f\in\mathcal{F}_R}\left|E[\phi_f] - \frac{1}{m}\sum_{i=1}^{m}\phi_f(z_i)\right|\right] \leq \widetilde{C}(3M + R)^2 \max\left\{(E[\phi_f])^{\frac{1-s}{2}} m^{-\frac{1}{2}}, m^{-\frac{1}{1+s}}\right\}. \quad (51)$$

Proof For arbitrary $\phi_f \in \mathcal{F}_R$, we have from (48) $|y_i \leq M$ and $\kappa \leq 1$ that

$$\|\phi_f\|_\infty \leq (3M + R)^2 =: B_1, \qquad E[\phi_f^2] \leq (3M + R)^2 E[\phi_f] =: \sigma_1^2.$$

Let $\bar{c} \geq 1$ be the smallest constant such that $\bar{c}c_s c'_s\sqrt{c} \geq 1$, that is,

$$B_1 = (3M+R)^2 \leq \bar{c}c_s c'_s\sqrt{c}(3M+R)^2 =: a.$$

Then, (47) implies

$$E[e_i(\mathcal{F}_R, \ell^2(D))] \leq ai^{-\frac{1}{2s}}.$$

Thus, it follows from Lemma 26 with $p = s$ that

$$E[Rad_D(\mathcal{F}_R, m)] \leq C'(3M+R)^2 \max\left\{(E[\phi_f])^{\frac{1-s}{2}} m^{-\frac{1}{2}}, m^{-\frac{1}{1+s}}\right\},$$

where $C' = \max\{C_s(\bar{c}c_s c'_s\sqrt{c})^s, C'_s(\bar{c}c_s c'_s\sqrt{c})^{\frac{2s}{1+s}}\}$. Based on Lemma 27, we then get

$$E\left[\sup_{\phi_f \in \mathcal{F}_R}\left|E[\phi_f] - \frac{1}{m}\sum_{i=1}^m \phi_f(z_i)\right|\right] \leq 2C'(3M+R)^2 \max\left\{(E[\phi_f])^{\frac{1-s}{2}} m^{-\frac{1}{2}}, m^{-\frac{1}{1+s}}\right\}.$$

This proves Lemma 28 with $\widetilde{C} = 2C'$. ■

For $\varepsilon > 0$, define

$$\mathcal{G}_{R,\varepsilon} := \left\{g_{\phi_f,\varepsilon} = \frac{E[\phi_f] - \phi_f}{E[\phi_f] + \varepsilon} : \phi_f \in \mathcal{F}_R\right\}. \tag{52}$$

Lemma 28 implies the following estimate.

Lemma 29 *If $|y_i| \leq M$, $\kappa \leq 1$ and Assumption 1 holds with some $c > 0$ and $0 < s < 1$, then for arbitrary $\varepsilon \geq \inf_{\phi_f \in \mathcal{F}_R} E[\phi_f]$, there exists a constant $\widetilde{C}_1$ depending only on c and s such that*

$$E\left[\sup_{g_{\phi_f,\varepsilon} \in \mathcal{G}_{R,\varepsilon}}\left|\frac{1}{m}\sum_{i=1}^m g_{\phi_f,\varepsilon}(z_i)\right|\right] \leq \widetilde{C}_1 \frac{(3M+R)^2}{\varepsilon} \max\left\{\varepsilon^{\frac{1-s}{2}} m^{-\frac{1}{2}}, m^{-\frac{1}{1+s}}\right\}.$$

Proof For arbitrary $\phi_f \in \mathcal{F}_R$, it follows from (48) that $E[\phi_f] = \mathcal{E}(f) - \mathcal{E}(f_\rho) \geq 0$. Then,

$$\sup_{\phi_f \in \mathcal{F}_R}\left|\frac{E[\phi_f] - \frac{1}{m}\sum_{i=1}^m \phi_f(z_i)}{E[\phi_f] + \varepsilon}\right| \leq \sup_{\phi_f \in \mathcal{F}_R, E[\phi_f] \leq \varepsilon} \frac{\left|E[\phi_f] - \frac{1}{m}\sum_{i=1}^m \phi_f(z_i)\right|}{\varepsilon}$$
$$+ \sum_{j=0}^{\infty} \sup_{\phi_f \in \mathcal{F}_R, 4^j\varepsilon \leq E[\phi_f] \leq 4^{j+1}\varepsilon} \frac{\left|E[\phi_f] - \frac{1}{m}\sum_{i=1}^m \phi_f(z_i)\right|}{4^j\varepsilon + \varepsilon},$$

where we used the convention $\sup \varnothing := 0$. Let $r \geq \inf_{\phi_f \in \mathcal{F}_R} E[\phi_f]$ be arbitrary real number. It follows from (51) that

$$E\left[\sup_{\phi_f \in \mathcal{F}_R, E[\phi_f] \leq r}\left|E[\phi_f] - \frac{1}{m}\sum_{i=1}^m \phi_f(z_i)\right|\right] \leq \widetilde{C}(3M+R)^2 \max\left\{r^{\frac{1-s}{2}} m^{-\frac{1}{2}}, m^{-\frac{1}{1+s}}\right\}.$$

Repeating the above inequality with $r = 4^j\varepsilon$ and $\varepsilon \geq \inf_{\phi_f \in \mathcal{F}_R} E[\phi_f]$ for $j = 0, 1, \ldots$, we get from the above two estimates that

$$\begin{aligned} & E\left[\sup_{\phi_f \in \mathcal{F}_R}\left|\frac{E[\phi_f] - \frac{1}{m}\sum_{i=1}^m \phi(z_i)}{E[\phi_f] + \varepsilon}\right|\right] \leq \frac{\widetilde{C}(3M+R)^2 \max\left\{\varepsilon^{\frac{1-s}{2}} m^{-\frac{1}{2}}, m^{-\frac{1}{1+s}}\right\}}{\varepsilon} \\ + & \sum_{j=0}^{\infty} \frac{\widetilde{C}(3M+R)^2 \max\left\{(4^{j+1}\varepsilon)^{\frac{1-s}{2}} m^{-\frac{1}{2}}, m^{-\frac{1}{1+s}}\right\}}{4^j\varepsilon + \varepsilon} \\ \leq & \frac{\widetilde{C}(3M+R)^2}{\varepsilon} \max\left\{\varepsilon^{\frac{1-s}{2}} m^{-\frac{1}{2}}, m^{-\frac{1}{1+s}}\right\} \left(1 + \sum_{j=0}^{\infty} \frac{4^{\frac{(1-s)(j+1)}{2}}}{4^j + 1}\right). \end{aligned}$$

Since

$$\sum_{j=0}^{\infty} \frac{4^{\frac{(1-s)(j+1)}{2}}}{4^j + 1} \leq 2^{1-s} \sum_{j=0}^{\infty} 2^{(-s-1)j} = \frac{2^{1-s}}{1 - 2^{-s-1}} \leq 4,$$

we get from (52) that

$$\begin{aligned} & E\left[\sup_{g_{\phi_f,\varepsilon} \in \mathcal{G}_{R,\varepsilon}}\left|\frac{1}{m}\sum_{i=1}^m g_{\phi_f,\varepsilon}(z_i)\right|\right] = E\left[\sup_{\phi_f \in \mathcal{F}_R}\left|\frac{E[\phi_f] - \frac{1}{m}\sum_{i=1}^m \phi_f(z_i)}{E[\phi_f] + \varepsilon}\right|\right] \\ \leq & 4\widetilde{C}\frac{(3M+R)^2}{\varepsilon} \max\left\{\varepsilon^{\frac{1-s}{2}} m^{-\frac{1}{2}}, m^{-\frac{1}{1+s}}\right\}. \end{aligned}$$

This completes the proof of Lemma 29 with $\widetilde{C}_1 := 4\widetilde{C}$. ■

Lemma 29 builds the estimate in expectation. To derive similar bound in probability, we need the following concentration inequality, which is a simplified version of Talagrand's inequality and can be found in (Steinwart and Christmann, 2008, Theorem 7.5, Lemma 7.6)

Lemma 30 *Let $B \geq 0$ and $\sigma \geq 0$ be constants such that $E[g^2] \leq \sigma^2$ and $\|g\|_\infty \leq B$ for all $g \in \mathcal{G}_{R,\varepsilon}$. Then, for all $\tau > 0$ and all $\gamma > 0$, we have*

$$\begin{aligned} & P\left(\left\{z \in \mathcal{Z}^m : \sup_{g_{\phi_f,\varepsilon} \in \mathcal{G}_{R,\varepsilon}}\left|\frac{1}{m}\sum_{i=1}^m g_{\phi_f,\varepsilon}(z_i)\right| \geq (1+\gamma) E\left[\sup_{g_{\phi_f,\varepsilon} \in \mathcal{G}_{R,\varepsilon}}\left|\frac{1}{m}\sum_{i=1}^m g_{\phi_f,\varepsilon}(z_i)\right|\right]\right.\right. \\ + & \left.\left. \sqrt{\frac{2\tau\sigma^2}{m}} + \left(\frac{2}{3} + \frac{1}{\gamma}\right)\frac{\tau B}{m}\right\}\right) \leq e^{-\tau}. \end{aligned} \tag{53}$$

Now, we are in a position to prove Theorem 14 by using Lemma 30 and lemma 29.

Proof of Theorem 14. For arbitrary $f \in B_{K,R}$, we have $E[\phi_f] = \mathcal{E}(f) - \mathcal{E}(f_\rho) \geq 0$ with $\phi_f = (y - f(x))^2 - (y - f_\rho(x))^2 \in \mathcal{F}_R$. Furthermore, $|y_i| \leq M$ and $\|f\|_\infty \leq \|f\|_K \leq R$ yield $\|E[\phi_f] - \phi_f\|_\infty \leq 2(R + 3M)^2$. For arbitrary $\varepsilon \geq \inf_{\phi_f \in \mathcal{F}_R} E[\phi_f]$ and $g_{\phi_f,\varepsilon} \in \mathcal{G}_{\phi_f,\varepsilon}$, there exists a $\phi_f \in \mathcal{F}_R$ such that $g_{\phi_f,\varepsilon} = \frac{E[\phi_f] - \phi_f}{E[\phi_f] + \varepsilon}$. Then, we get

$$\|g_{\phi_f,\varepsilon}\|_\infty \leq \frac{2(3M+R)^2}{\varepsilon} =: B, \tag{54}$$

and

$$E[g^2_{\phi_f,\varepsilon}] \le \frac{E[\phi_f^2]}{(E[\phi_f+\varepsilon)^2} \le \frac{(3M+R)^2 E[\phi_f]}{(E[\phi_f+\varepsilon)^2} \le \frac{(3M+R)^2}{\varepsilon}. \tag{55}$$

Then Lemma 30 with $\gamma = 1$ and $\varepsilon \ge \inf_{\phi_f\in\mathcal{F}_R} E[\phi_f]$, Lemma 29 with $\varepsilon \ge \inf_{\phi_f\in\mathcal{F}_R} E[\phi_f]$, (54) and (55) that with confidence at least $1-e^{-\tau}$, there holds

$$\begin{aligned}
&\sup_{\phi_f\in\mathcal{F}_R}\left|\frac{E[\phi_f]-\frac{1}{m}\sum_{i=1}^m \phi_f(z_i)}{E[\phi_f]+\varepsilon}\right| = \sup_{g_{\phi_f,\varepsilon}\in\mathcal{G}_{R,\varepsilon}}\left|\frac{1}{m}\sum_{i=1}^m g_{\phi_f,\varepsilon}(z_i)\right| \\
\le\ & 2E\left[\sup_{g_{\phi_f,\varepsilon}\in\mathcal{G}_{R,\varepsilon}}\left|\frac{1}{m}\sum_{i=1}^m g_{\phi_f,\varepsilon}(z_i)\right|\right] + \sqrt{\frac{2\tau(3M+R)^2}{m\varepsilon}} + \frac{10(3M+R)^2\tau}{3m\varepsilon} \\
\le\ & 2\widetilde{C}_1\frac{(3M+R)^2}{\varepsilon}\max\left\{\varepsilon^{\frac{1-s}{2}}m^{-\frac{1}{2}}, m^{-\frac{1}{1+s}}\right\} + \sqrt{\frac{2\tau(3M+R)^2}{m\varepsilon}} + \frac{10(3M+R)^2\tau}{3m\varepsilon}.
\end{aligned}$$

For arbitrary $f \in B_{K,R}$, set $\tau = \log\frac{1}{\delta}$ and $\varepsilon = \mathcal{E}(f) - \mathcal{E}(f_\rho) \ge \inf_{\phi_f\in\mathcal{F}_R} E[\phi_f]$. It follows from $\mathcal{E}(f) - \mathcal{E}_D(f) = E[\phi_f] - \frac{1}{m}\sum_{i=1}^m \phi_f(z_i)$ that , with confidence $1-\delta$, there holds

$$\begin{aligned}
&|\mathcal{E}(f) - \mathcal{E}(f_\rho) + \mathcal{E}_D(f_\rho) - \mathcal{E}_D(f_\rho)| \le \sqrt{\frac{8(3M+R)^2(\mathcal{E}(f)-\mathcal{E}(f_\rho))\log\frac{1}{\delta}}{m}} \\
+\ & \frac{20(3M+R)^2\log\frac{1}{\delta}}{3m} + 4\widetilde{C}_1(3M+R)^2\max\left\{(\mathcal{E}(f)-\mathcal{E}(f_\rho))^{\frac{1-s}{2}}m^{-\frac{1}{2}}, m^{-\frac{1}{1+s}}\right\} \\
\le\ & \frac{1}{2}(\mathcal{E}(f)-\mathcal{E}(f_\rho)) + \frac{32(3M+R)^2\log\frac{1}{\delta}}{3m} \\
+\ & 4\widetilde{C}_1(3M+R)^2\max\left\{(\mathcal{E}(f)-\mathcal{E}(f_\rho))^{\frac{1-s}{2}}m^{-\frac{1}{2}}, m^{-\frac{1}{1+s}}\right\},
\end{aligned}$$

where we used the element inequality $\sqrt{ab} \le \frac{1}{2}(a+b)$ for $a, b > 0$ in the last inequality. This completes the proof of Theorem 14 with $\widetilde{C} = 4\widetilde{C}$. ■

Acknowledgments

The research was partially supported by the Major Key Project of PCL under Grant PCL2024A06 and the National Natural Science Foundation of China (Grant Nos.12371513, 62276209). Part of the work of Xin Guo was done when he worked at The Hong Kong Polytechnic University and supported partially by the Research Grants Council of Hong Kong [Project No. PolyU 15305018]. Xin Guo acknowledges funding from the ARC grant: DP230100905.

References

A. Bagirov, C. Clausen, and M. Kohler. An L_2 boosting algorithm for estimation of a regression function. *IEEE. Trans. Inf. Theory.* 56: 1417-1429, 2010.

A. R. Barron, A. Cohen, W. Dahmen, and R. A. DeVore. Approximation and learning by greedy algorithms. *Ann. Statist.* 36: 64-94, 2008.

F. Bauer, S. Pereverzev, and L. Rosasco. On regularization algorithms in learning theory. *J. Complex.*, 23(1):52–72, 2007.

M. Belkin, P. Niyogi, and V. Sindhwani. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *J. Mach. Learn. Res.*, 7: 2399-2434, 2006.

P. Bickel, Y. Ritov, and A. Zakai. Some theory for generalized boosting algorithms. *J. Mach. Learn. Res.* 7: 705-732, 2006.

P. Bühlmann and B. Yu. Boosting with the L_2 loss: regression and classification. *J. Amer. Satis. Assoc.*, 98: 324-339, 2003.

G. Blanchard and N. Krämer. Convergence rates for kernel conjugate gradient for random design regression. *Anal. Appl.*, 14: 763-794, 2016.

G. Blanchard, P. Mathé, and N. Mück. Lepskii principle in supervised learning. arXiv preprint arXiv:1905.10764.

L. Breiman, J. Friedman, C. Stone, and R. Olshen. Classification and Regression Trees. *CRC press*, 1984.

A. Caponnetto and E. DeVito. Optimal rates for the regularized least squares algorithm. *Found. Comput. Math.*, 7: 331-368, 2007.

A. Caponnetto and Y. Yao. Cross-validation based adaptation for regularization operators in learning theory. *Anal. Appl.*, 8: 161-183, 2010.

A. Celisse and M. Wahl. Analyzing the discrepancy principle for kernelized spectral filter learning algorithms. *J. Mach. Learn. Res.*, 22: 1-59, 2022.

X. Chang, S. B. Lin, and D. X. Zhou. Distributed semi-supervised learning with kernel ridge regression. *J. Mach. Learn. Res.*, 18: 1-22, 2017.

F. Cucker and D. X. Zhou. *Learning Theory: An Approximation Theory Viewpoint*. Cambridge University Press, 2007.

R. DeVore and V. Temlyakov. Some remarks on greedy algorithms. *Adv. Comput. Math.*, 5: 173-187, 1996.

N. Duffy and D. Helmbold. Boosting methods for regression. *Mach. Learn.* 47: 153-200, 2002.

B. Efron, T. Hastie, I. Johnstone, and R. Tibsirani. Least angle regressionn. *Ann. Statist.*, 32: 407-451, 2004.

J. Ehrlinger and H. Ishwaran. Characterizing L_2 boosting. *Ann. Statist.*, 40: 1074-1101, 2012.

T. Evgeniou, M. Pontil, and T. Poggio. Regularization networks and support vector machines. *Adv. Comput. Math.*, 13: 1-50, 2000.

Y. Feng, S. G. Lv, H. Hang, and J. A. Suykens. Kernelized elastic net regularization: generalization bounds, and sparse recovery. *Neural Comput.*, 28: 525-562, 2016.

Y. Freund. Boosting a weak learning algorithm by majority. *Inform. & Comput.*, 121: 256-285, 1995.

J. H. Friedman. Greedy function approximation: a gradient boosting machine. *Ann. Stat.*, 29: 1189-1232, 2001.

L. L. Gerfo, L. Rosasco, F. Odone, E. De Vito, and A. Verri. Spectral algorithms for supervised learning. *Neural Comput.*, 20: 1873-1897, 2008.

A. Grittens and M. W. Mahoney. Revisitng the Nyström method for improved large scale machine learning. *J. Mach. Learn. Res.*, 17: 1-65, 2016.

Z. C. Guo and L. Shi. Learning with coefficient-based regularization and ℓ_1-penalty. *Adv. Comput. Math.*, 39: 493-510, 2013.

Z. C. Guo, S. B. Lin, and D. X. Zhou. Learning theory of distributed spectral algorithms. *Inverse Probl.*, 33: 074009, 2017.

Z. C. Guo, D. H. Xiang, X. Guo, and D. X. Zhou. Thresholded spectral algorithms for sparse approximations. *Anal. Appl.*, 15: 433-455, 2017.

Z. C. Guo, L. Shi, and Q. Wu. Learning theory of distributed regression with bias corrected regularization kernel network. *J. Mach. Learn. Res.*, 18(1): 4237-4261, 2017.

L. Györfy, M. Kohler, A. Krzyzak, and H. Walk. *A Distribution-Free Theory of Nonparametric Regression.* Springer, Berlin, 2002.

T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning.* Springer, New York, 2001.

T. Hastie, J. Taylor, R. Tibshirani, and G. Walther. Forward stagewise regression and the monotone lasso. *Elec. J. Statis.* 1: 1-29, 2007.

S. B. Lin, X. Guo, and D. X. Zhou. Distributed learning with regularized least squares. *J. Mach. Learn. Res.*, 18: 1-31, 2017.

S. B. Lin, Y. Lei, and D. X. Zhou. Boosted kernel ridge regression: optimal learning rates and early stopping. *J. Mach. Learn. Res.*, 20(46): 1-36, 2019.

S. Lin, J. Zeng, F. Fang, and Z. Xu. Learning rates of ℓ_q coefficient regularization learning with Gaussian kernel. *Neural Comput.*, 26(10): 2350-2378, 2014.

S. B. Lin and D. X. Zhou. Distributed kernel-based gradient descent algorithms. *Constr. Approx.*, 47:249–276, 2018.

S. B. Lin and D. X. Zhou. Optimal learning rates for kernel partial least squares. *J. Fourier Anal. Appl.*, 24: 908-933, 2018.

E. Livshits. Lower bounds for the rate of convergence of greedy algorithms. *Izvestiya: Math.,* 73: 1197-1215, 2009.

S. Lu, P. Mathé, and S. Pereverzyev Jr. Analysis of regularized Nystr?m subsampling for regression functions of low smoothness. *Analysis and Applications*, 17(06): 931-946, 2019.

S. Lu, P. Mathé, and S. Pereverzyev. Balancing principle in supervised learning for a general regularization scheme. *Appl. Comput. Harmon. Anal.*, 48(1): 123-148, 2020.

M. Meister and I. Steinwart. Optimal learning rates for localized SVMs. *J. Mach. Learn. Res.*, 17: 1-44, 2016.

I. Mukherjee, C. Rudin, and R. E. Schapire. The rate of convergence of AdaBoost. *J. Mach. Learn. Res.*, 14: 2315-2347, 2013.

W. Nash, T. Sellers, S. Talbot, A. Cawthorn, and W. Ford. The population biology of abalone (haliotis species) in tasmania. i. blacklip abalone (h. rubra) from the north coast and islands of bass strait. *Sea Fisheries Division, Tech. Rep.*, no. 48, 1994.

G. Petrova. Rescaled pure greedy algorithm for Hilbert and Banach spaces. *Appl. Comput. Harmonic Anal.*, 41: 852-866, 2016.

G. Raskutti, M. Wainwright, and B. Yu. Early stopping and non-parametric regression: an optimal data-dependent stopping rule. *J. Mach. Learn. Res.*, 15: 335-366, 2014.

A. Rudi, R. Camoriano, and L. Rosasco. Less is more: Nyström computational regularization. *Advances in Neural Information Processing Systems*, 1657-1665, 2015.

S. Shalev-Shwartz, N. Srebro, and T. Zhang. Trading accuracy for sparsity in optimization problems with sparsity constraints. *SIAM J. Optimiz.*, 20: 2807-2832, 2010.

L. Shi, Y. L. Feng, and D. X. Zhou. Concentration estimates for learning with l_1-regularizer and data dependent hypothesis spaces. *Appl. Comput. Harmonic Anal.*, 31: 286-302, 2011.

L. Shi. Learning theory estimates for coefficient-based regularized regression. *Appl. Comput. Harmon. Anal.*, 34: 252-265, 2013.

I. Steinwart and A. Christmann. *Support Vector Machines*. Springer, New York, 2008.

I. Steinwart, D. Hush, and C. Scovel. Optimal rates for regularized least squares regression. In S. Dasgupta and A. Klivan, editors, *Annual Conference on Learning Theory*, pages 79-93, 2009.

I. Steinwart and C. Scovel. Mercer's theorem on general domains: On the interaction between measures, kernels, and RKHSs. *Constructive Approximation*, 35: 363-417, 2012.

V. Temlyakov. Relaxation in greedy approximation. *Constr. Approx.*, 28: 1-25, 2008.

V. Temlyakov. Greedy approximation in convex optimization. *Constr. Approx.*, 41: 269-296, 2015.

G. Wahba. *Spline models for observational data*. Society for Industrial and Applied Mathematics, 1990.

G. Wang, D. Y. Yeung, and F. H. Lochovsky. The kernel path in kernelized LASSO. *Artificial Intelligence and Statistics*, 580-587, 2007.

Y. Wang, X. Liao, and S. Lin. Rescaled boosting in classification. *IEEE Trans. Neural Netw. & Learn. Syst.*, 30: 2598-2610, 2019.

Y. Wei, F. Yang, and M. J. Wainwright. Early stopping for kernel boosting algorithms: A general analysis with localized complexities. *IEEE Trans. Inf. Theory*, 65: 6685-6703, 2019.

C. K. I. Williams and M. Seeger. Using the Nyström method to speed up kernel machines. *Proceedings of the 13th International Conference on Neural Information Processing Systems*, pages 661-667, 2000.

L. Xu, S. Lin, Y. Wang, and Z. Xu. Shrinkage degree in L_2-rescale boosting for regression. *IEEE Trans. Neural Netw. & Learn. Syst.*, 28: 1851-1864, 2017.

Y. Yao, L. Rosasco, and A. Caponnetto. On early stopping in gradient descent learning. *Constr. Approx.*, 26: 289-315, 2007.

Modeling of strength of high performance concrete using artificial neural networks. *Cem Concr Res.*, 28: 1797-1808, 1998.

T. Zhang and B. Yu. Boosting with early stopping: convergence and consistency. *Ann. Statis.* 33: 1538-1579, 2005.

Y. C. Zhang, J. Duchi, and M. Wainwright. Divide and conquer kernel ridge regression: A distributed algorithm with minimax optimal rates. *J. Mach. Learn. Res.*, 16: 3299-3340, 2015.

H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *J. Royal Statist. Soc.: Series B*, 67 (2): 301-320, 2005.